J Math Imaging Vis
DOI 10.1007/s10851-012-0399-y

3D Deformable Super-Resolution for Multi-Camera 3D Face Scanning

Karima Ouji · Mohsen Ardabilian · Liming Chen · Faouzi Ghorbel

Abstract Low-cost and high-accuracy 3D face measurement is becoming increasingly important in many computer vision applications including face recognition, facial animation, games, orthodontics and aesthetic surgery. In most cases fringe projection based systems are used to overcome the relatively uniform appearance of skin. These systems employ a structured light camera/projector device and require explicit user cooperation and controlled lighting conditions. In this paper, we propose a 3D acquisition solution with a 3D space-time non-rigid super-resolution capability, using three calibrated cameras coupled with a non calibrated projector device, which is particularly suited to 3D face scanning, i.e. rapid, easily movable and robust to ambient lighting variation. The proposed solution is a hybrid stereovision and phase-shifting approach, using two shifted patterns and a texture image, which not only takes advantage of stereovision and structured light, but also overcomes their weaknesses. The super-resolution scheme involves a shape+texture 3D non-rigid registration for 3D artifacts correction in the presence of small non-rigid deformations as facial expressions.

Keywords Stereovision · Phase-shifting · Space-time · Multi-camera · Super-resolution · Non-rigid registration

1 Introduction

Recently, low-cost and high-accuracy 3D face measurement systems are increasingly demanded for many applications like face recognition, facial animation, games, orthodontics and aesthetic surgery. In most cases fringe projection based systems are used to overcome the relatively uniform appearance of skin. These systems employ a structured light camera/projector device and require explicit user cooperation and controlled lighting conditions [22, 24]. Depth information is recovered by decoding patterns of the projected structured light. Current solutions mostly utilize more than three phase-shifted sinusoïdal patterns to recover the depth information, thus impacting the acquisition delay; they further require projector-camera calibration whose accuracy is crucial for phase to depth estimation step; and finally, they also need an unwrapping stage which is sensitive to ambient light, especially when the number of patterns decreases [23]. An alternative to projector-camera systems consists of recovering depth information by stereovision using a multi-camera system as proposed in [5, 24]. A stereo matching step finds correspondence between stereo images and the 3D information is obtained by optical triangulation [13, 24]. However, the model computed in this way generally is quite sparse. To upsample and denoise depth images, researchers looked into super-resolution techniques. Kil et al. [10] applied super-resolution for laser triangulation scanners by regular resampling from aligned scan points with associated

This research is supported in part by the ANR project FAR3D under the grant ANR-07-SESU-003.

K. Ouji (✉) · M. Ardabilian · L. Chen
LIRIS, Lyon Research Center for Images and Intelligent Information Systems, Ecole Centrale de Lyon, 36, av. Guy de Collongue, 69134 Ecully, France
e-mail: karima.ouji@ec-lyon.fr

M. Ardabilian
e-mail: mohsen.ardabilian@ec-lyon.fr

L. Chen
e-mail: liming.chen@ec-lyon.fr

F. Ghorbel
University of Manouba, 2010 Manouba, Tunisia
e-mail: faouzi.ghorbel@ensi.rnu.tn

Published online: 10 November 2012

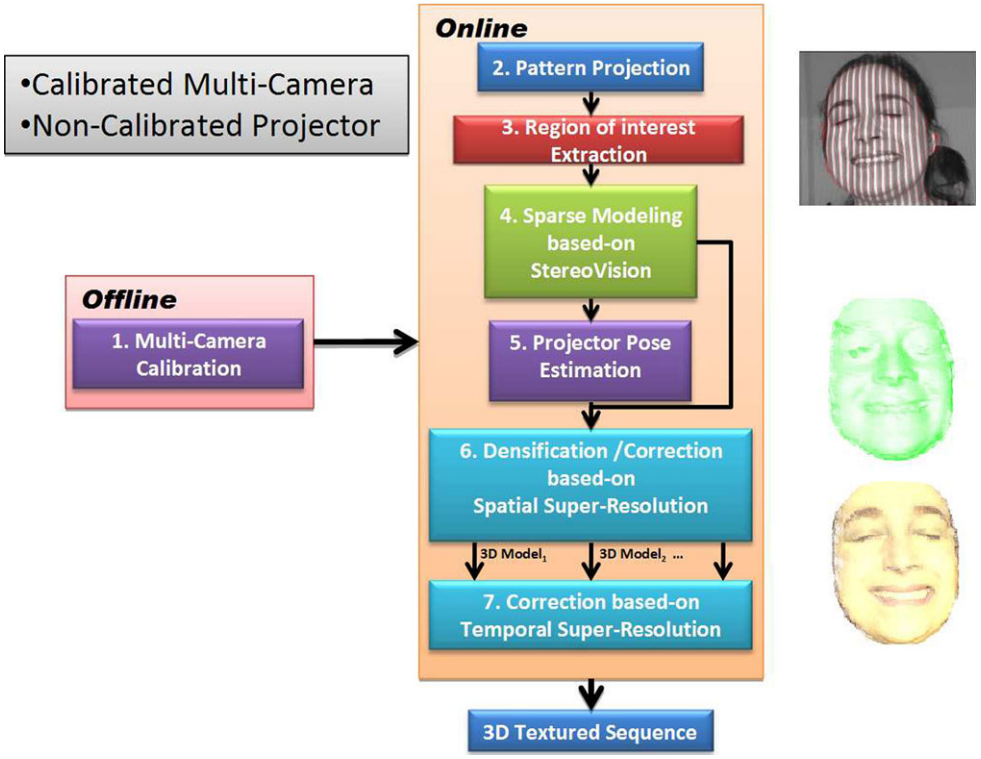


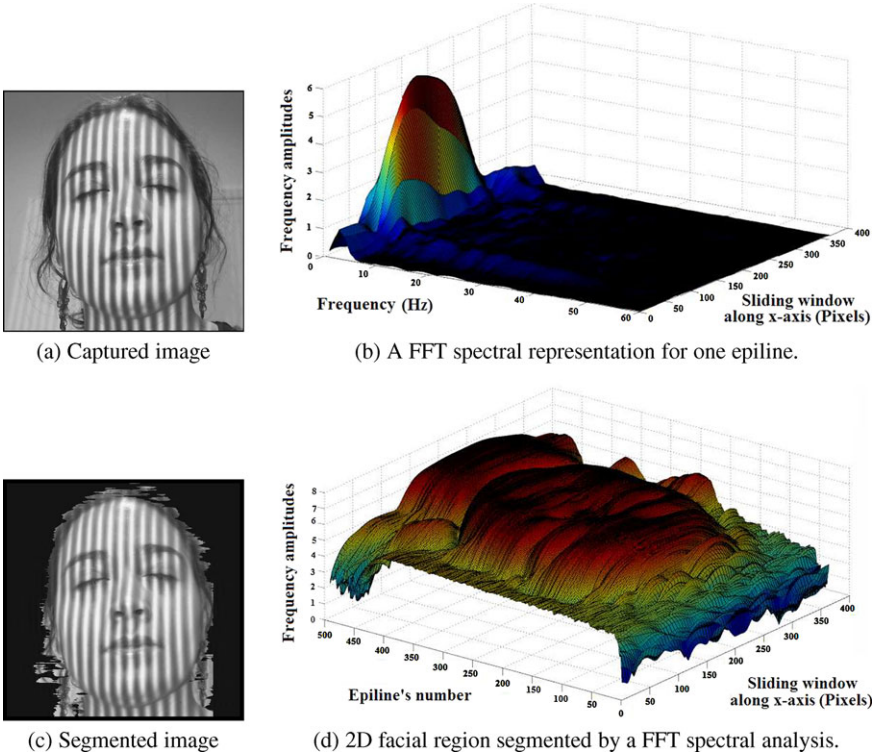
Fig. 1 3D Sequence Acquisition Framework of a textured moving face.

Gaussian location uncertainty. Super-resolution was especially proposed for time-of-flight cameras which have very low data quality and a very high random noise.

Rajagopalan et al. [14] proposed a Markov-Random-Field scheme which formulates the upsampled 3D geometry as the most likely surface given several low resolution measurements. Schuon et al. [16] suggested LidarBoost, a 3D depth superresolution method based on an energy minimization framework tailored to the specific characteristics of flash lidars. The framework jointly employs a data fidelity term and a geometry prior term enforcing similarity between the input and output images, and a bilateral regularization term for edge-preserving smoothness. Cui et al. [6] carried out a 3D super-resolution processing pipeline for the 3D Kinect scanner to enhance its reliability. The authors applied a loop closure alignment to deal with the rigid and non-rigid deformations.

In this paper, we propose a 3D acquisition solution with a 3D space-time and non-rigid super-resolution capability, using three calibrated cameras coupled with a non calibrated projector device, which is particularly suited to 3D face scanning, i.e. rapid, easily movable and robust to ambient lighting variation. The proposed solution is a hybrid stereovision and phase-shifting approach which not only takes advantage of stereovision and structured light, but also overcomes their weaknesses. Figure 1 presents our 3D face scanning scheme. According to our method, first an automatic primitives sampling is performed from stereo-matching to provide a 3D facial sparse model with a fringe-based resolution and a subpixel precision. Second, an intra-fringe phase estimation densify the 3D sparse model using the two sinusoïdal fringe images and a texture image, independently from the left, middle and right cameras. The left, middle and right 3D dense models are merged to produce the final 3D model which constitutes a spatial super-resolution. Also, we propose to carry out a shape+texture temporal super-resolution to correct the 3D information and to complete the 3D scanned view. Our temporal super-resolution scheme is based on a non-rigid registration step to deal with facial expression deformations. In contrast to conventional structured-light methods, the use of stereo in the first stage of the approach replaces the phase unwrapping stage. Also, it does not require a camera-projector off-line calibration which constitutes a tedious and expensive task. Moreover, our approach is applied only to the region of interest which reduces the total processing time.

Mainly, two practical applications can be tackled with our 3D scanning approach. First, it can be useful for facial animation and identification. Second, it finds its application in the medical field for orthodontics and reconstructive surgery. For orthodontics, the scanner should be easily movable to capture plaster dental arch models. A camera-projector system has the problem of measurement shadow caused by projector and occlusion caused by the camera. Therefore, using two cameras with a mobile projector as we propose solves these problems. Also, our scanner can capture facial motion and help facial mimics study after a maxillofacial surgery or a rhinoplasty. Section 2 details



(a) Captured image

(b) A FFT spectral representation for one epiline.

(c) Segmented image

(d) 2D facial region segmented by a FFT spectral analysis.

Fig. 2 Pattern-based face localization

the primitives sampling to generate the 3D sparse model. In Sect. 3, we high-light the spatial super-resolution from the three calibrated cameras. Section 4 explains how the 3D non-rigid temporal super-resolution is carried out. Section 5 discusses the experimental results and Sect. 7 concludes the paper.

2 Primitives Sampling for 3D Sparse Model Generation

First, an offline strong stereo calibration computes the intrinsic and extrinsic parameters of the cameras, estimates the tangential and radial distortion parameters, and provides the epipolar geometry as proposed in [21]. In online process, two π-shifted sinusoïdal patterns and a third white pattern are projected onto the face. Three sets of left, middle and right images are captured, undistorted and rectified. The proposed model is defined by the system of Eqs. (1). It constitutes a variant of the mathematic model proposed in [23].

$$\begin{aligned} I_p(s,t) &= I_b(s,t) + I_a(s,t)\cdot\sin(\phi(s,t)),\\ I_n(s,t) &= I_b(s,t) + I_a(s,t)\cdot\sin(\phi(s,t)+\pi),\\ I_t(s,t) &= I_b(s,t) + I_a(s,t). \end{aligned} \tag{1}$$

At time t , $I_p(s,t)$, $I_n(s,t)$, $I_t(s,t)$ constitute the intensity term of the pixel s on respectively the positive image, the negative one and the texture one. $I_b(s,t)$ represents the texture information and the lighting effect. $\phi(s,t)$ is the local phase defined at each pixel s. Solving (1), $I_b(s,t)$ is computed as the average intensity of $I_p(s,t)$ and $I_n(s,t)$. $I_a(s,t)$ is then computed from the third equation of the system (1) and $\phi(s,t)$ is estimated by Eq. (2).

$$\phi(s,t) = \arcsin\left[\frac{I_p(s,t) - I_n(s,t)}{2\cdot I_t(s,t) - I_p(s,t) - I_n(s,t)}\right]. \tag{2}$$

Also, we suggest an automatic region-of interest localization to reduce the total processing time. The idea is to benefit from the contrast variation and carry out a spectral analysis to localize the low frequencies on captured images. First, we compute FFT on a sliding window for each epiline which provides for each pixel a 2D curve of FFT frequency amplitudes. A 3D spectral distribution is obtained which highlights the facial region for the current epiline as shown in Fig. 2.b. We propose to keep only pixels belonging to this highlighted region. Thus, for each pixel in the epiline, we consider a weighted sum of only the low-frequency amplitudes and we apply an adequate thresholding to obtain the region-of-interest as illustrated by Fig. 2.d.

Finally, the sparse 3D model is generated through a stereovision scenario. It is formed by the primitives situated on the fringe change-over which is the intersection of the sinusoïdal component of the positive image and the second π-shifted sinusoïdal component of the negative one [13]. Therefore, the primitives localization has a sub-pixel precision. Corresponding multi-camera primitives necessarily have the same Y-coordinate in the rectified images. Thus,

stereo matching problem is resolved in each epiline separately using Dynamic Programming. The 3D sparse point cloud is then recovered by computing the intersection of optical rays coming from the pair of matched features. When projecting vertical fringes, the video projector can be considered as a vertical adjacent sources of light. Such a consideration provides for each epiline a light source point O_{Prj} situated on the corresponding epipolar plane. The sparse 3D model is a serie of adjacent 3D vertical curves obtained by the fringes intersection of the positive and the negative images. Each curve describes the profile of a projected vertical fringe distorted on the 3D facial surface. We propose to estimate the 3D plane containing each distorted 3D curve separately. As a result, the light source vertical axis of the projector is defined as the intersection of all the computed 3D planes. This estimation can be performed either as an offline or online process unlike conventional phase-shifting approaches where the projector is calibrated on offline and cannot change its position when scanning the object.

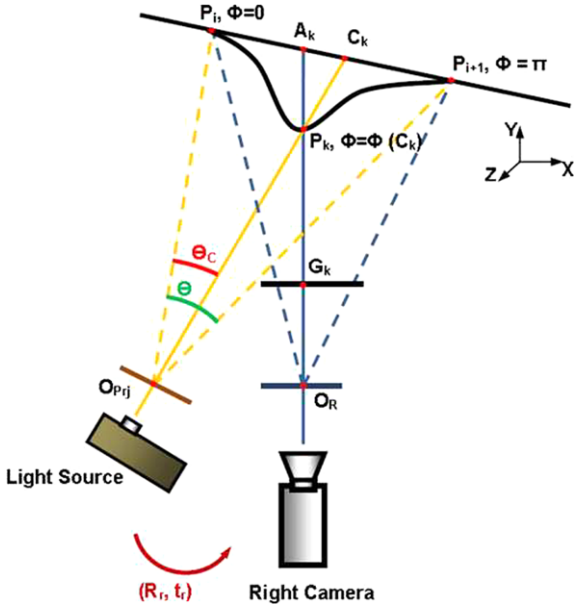


Fig. 3 Intra-fringe 3D information retrieval scheme

3 3D Multi-Camera Spatial Super-Resolution

We need to find the 3D coordinates for each pixel situated between two successive fringes in either left, middle or right camera images to participate separately on the 3D model elaboration. Thus, a 3D point cloud is obtained from each camera set of images. The spatial super-resolution consists of merging the left, middle and right 3D point clouds. A phase-shifting analysis allows an estimation of the 3D coordinates of each pixel separately. Conventional phase-shifting techniques estimates the local phase in $[0..2\pi]$ for each pixel on the captured image. Local phases are defined as wrapped phases. Absolute phases are obtained by phase unwrapping. In the proposed approach, the sparse model lets us retrieve 3D intra-fringe information from wrapped phases directly. In fact, each point P_i in the sparse model constitutes a reference point for all pixels situated between P_i and its next neighbor P_{i+1} on the same epiline of the sparse model. For a pixel P_k situated between $P_i(X_i, Y_i, Z_i)$ and $P_{i+1}(X_{i+1}, Y_{i+1}, Z_{i+1})$, we compute its local phase value ϕ_k using Eq. (2). The phase value of P_i is $\phi_i = 0$ and the phase value of P_{i+1} is $\phi_{i+1} = \pi$.

The phase ϕ_k which belongs to $[0..\pi]$ has monotonous variation if $[P_i P_{i+1}]$ constitutes a straight line on the 3D model. When $[P_i P_{i+1}]$ represents a curve on the 3D model, the function ϕ_k describes the depth variation inside $[P_i P_{i+1}]$. Therefore, the 3D coordinates $(X(\phi_k), Y(\phi_k), Z(\phi_k))$ of P_k corresponding to the pixel point G_k are computed by a geometric reconstruction as shown in Fig. 3.

The 3D intra-fringe coordinates computation is carried out for each epiline i separately. An epipolar plane is defined for each epiline and contains the optical centers O_L, O_M and O_R of respectively left, middle and right cameras and all 3D points situated on the current epiline i. Each 3D point P_k is characterized by its own phase value $\phi(P_k)$. The light ray coming from the light source into the 3D point P_k intersects the segment $[P_i P_{i+1}]$ in a 3D point C_k having the same phase value $\phi(C_k) = \phi(P_k)$ as P_k. To localize C_k, we need to find the distance $P_i C_k$. This distance is computed by applying the sine law in the triangle $O_{Prj} P_i C_k$ as described in Eq. (3).

$$\frac{P_i C_k}{\sin(\theta_C)} = \frac{O_{Prj} P_i}{\sin(\pi - (\theta_C + \alpha))}. \tag{3}$$

The distance $O_{Prj} P_i$ and the angle α between $(O_{Prj} P_i)$ and $(P_i P_{i+1})$ are known. Also, the angle θ between $(O_{Prj} P_i)$ and $(O_{Prj} P_{i+1})$ is known. Thus, the angle θ_C is defined by Eq. (4). After localizing C_k, the 3D point P_k is identified as the intersection between $(O_R G_k)$ and $(O_{Prj} C_k)$.

$$\theta_C = \frac{\pi}{\theta} \cdot \phi(C_k). \tag{4}$$

Conventional super-resolution techniques carry out a registration step between low-resolution data, a fusion step and a deblurring step. Here, the phase-shifting analysis provides a matched left, middle and right point clouds since their 3D coordinates are computed based on the same 3D sparse point cloud. Also, left, middle and right point clouds present homogeneous 3D data and need only to be merged to retrieve the high-resolution 3D point cloud.

4 3D Non-rigid Temporal Super-Resolution

We propose to perform a 3D temporal super-resolution to correct the 3D information provided by the spatial super-resolution and to deal with 3D artifacts caused by either an expression variation, an occlusion or even a facial surface reflectance. First, our temporal super-resolution approach performs a non-rigid shape+texture registration for each couple of successive 3D point sets M_{t-1} and M_t at each moment t. Once registered, the 3D point sets M_{t-1} and M_t and also their corresponding 2D texture images are used as a low resolution data to create a high resolution 3D point set and its corresponding texture.

4.1 Shape+Texture Based Non-rigid Registration

The 3D non-rigid registration problem is formulated as a maximum-likelihood estimation problem since the deformation between two successive 3D faces is non rigid in general. We carry out a shape+texture registration based on the 3D non-rigid CPD registration algorithm (Coherent Point Drift) proposed in [12] to match two successive 3D frames M_{t-1} and M_t by deforming M_{t-1}. Our algorithm considers the alignment of two textured point sets source M_{src} and destination M_{dst} as a probability density estimation problem and apply an iterative deformation of M_{src} to minimize its spatial deviation with M_{dst}. N_{src} is the number of textured points of M_{src} and $M_{src} = \{s_n | n = 1, \ldots, N_{src}\}$. N_{dst} constitutes the number of textured points of M_{dst} and $M_{dst} = \{d_n | n = 1, \ldots, N_{dst}\}$. Each textured point $P \in M_{src} \cup M_{dst}$ is a 1×6 vector which concatenates shape and texture information $P(XYZRGB)$.

The algorithm suggests first of all to represent each textured point s_n of M_{src} by a centroid $Ctroid(s_n)$ of a Gaussian mixture model GMM, $Ctroid(s_n)$ being a multi-variate Gaussian centered on s_n. Thereby, the whole point set M_{src} can be considered as a Gaussian Mixture Model characterized by the probability density function $p(x)$ as defined by Eq. (5).

$$\begin{aligned} p(x) &= \sum_{v=1}^{N_{src}+1} P(v)p(x|v), \\ p(x|v) &= \frac{1}{(2\pi\sigma^2)^3} \exp^{-\frac{\|x-s_v\|^2}{2\sigma^2}}. \end{aligned} \tag{5}$$

Also, a uniform distribution $p(x|N_{src}+1)$ is added to the mixture model to account for noise and outliers, $p(x|N_{src}+1) = \frac{1}{N_{dst}}$. We use equal isotropic covariances σ^2 and equal membership probabilities $P(v) = \frac{1}{N_{src}}$ for all GMM components ($v = 1, \ldots, N_{src}$). Denoting the weight of the uniform distribution as w, $0 \leq w \leq 1$, the mixture model takes the form:

$$p(x) = w\frac{1}{N_{dst}} + (1-w)\sum_{v=1}^{N_{src}} \frac{1}{N_{src}} p(x|v). \tag{6}$$

Core to this method is to fit the GMM centroids representing M_{src} to the point set M_{dst} and to force the GMM centroids to move coherently as a group to preserve the topological structure of the point sets [11]. The GMM centroid locations are reparameterized by a set of non-rigid parameters θ and estimate them by maximizing the likelihood or, equivalently, by minimizing the negative log-likelihood function $E(\theta, \sigma^2)$ defined by Eq. (7).

$$E(\theta, \sigma^2) = -\sum_{u=1}^{N_{dst}} \log \sum_{v=1}^{N_{src}+1} P(v)p(d_u|v). \tag{7}$$

The correspondence probability between two points s_v and d_u is defined as the posterior probability of the GMM centroid given the data point: $P(v|d_u) = P(v)p(d_u|v)/p(d_u)$ and we use the *EM* algorithm [4, 7] to find θ and σ^2. The idea of *EM* is first to guess the values of parameters ("old" parameter values) and then use the Bayes theorem to compute a posteriori probability distributions $P^{old}(v|d_u)$ of mixture components, which is the expectation or E-step of the algorithm. The "new" parameter values are then found by minimizing the expectation of the complete negative log-likelihood function [4] with respect to the "new" parameters, which is called the maximization or M-step of the algorithm. The Q function, called the objective function, is also an upper bound of the negative log-likelihood function((8)).

$$Q = -\sum_{u=1}^{N_{dst}} \sum_{v=1}^{N_{src}+1} P^{old}(v|d_u) \log(P^{new}(v)P^{new}(d_u|v)). \tag{8}$$

The *EM* algorithm proceeds by alternating between E- and M-steps until convergence. Ignoring the constants independent of θ and σ^2, we rewrite Eq. (8) as Eq. (9) where $N_{dst,p} = \sum_{u=1}^{N_{dst}} \sum_{v=1}^{N_{src}} P^{old}(v|d_u) \leq N_{src}$ (with $N_{dst} = N_{dst,p}$ only if $w = 0$). $\tau(s_v, \theta)$ denotes Transformation τ applied to s_v regarding to the set of the non-rigid transformation parameters θ.

$$\begin{aligned} Q(\theta, \sigma^2) = &-\frac{1}{\sigma^2} \sum_{u=1}^{N_{dst}} \sum_{v=1}^{N_{src}} P^{old}(v|d_u) \|d_u - \tau(s_v, \theta)\|^2 \\ &+ \frac{3N_{dst,p}}{2} \log(\sigma^2). \end{aligned} \tag{9}$$

P^{old} denotes the posterior probabilities of GMM components calculated using the previous parameter values as de-

scribed by Eq. (10).

$$P^{old}(v|d_u) = \frac{\exp^{-\frac{1}{2}\|\frac{d_u-\tau(s_v,\theta^{old})}{\sigma^{old}}\|^2}}{\sum_{k=1}^{N_{src}} \exp^{-\frac{1}{2}\|\frac{d_u-\tau(s_k,\theta^{old})}{\sigma^{old}}\|^2} + C} \tag{10}$$

where $C = (2\pi\sigma^2)^3 \frac{w}{1-w} \frac{N_{src}}{N_{dst}}$. Minimizing the function Q, we necessarily decrease the negative log-likelihood function E unless it is already at a local minimum. In order to deal with the non-rigid problem, Tikhonov regularization framework is used [11, 17]. The transformation τ is defined as the initial position plus a displacement function V, $\tau(M_{src}, V) = M_{src} + V$. The displacement function V is estimated using variational calculus and the norm of V is regularized to enforce the smoothness of the deformation [11].

4.2 Fusion and Deblurring Steps

we propose a fusion and deblurring approach based on the 2D super-resolution technique proposed in [8, 16]. The 3D model M_t cannot be represented by only one 2D disparity image since the points situated on the fringe change-over have sub-pixel precision. Also, the left, middle and right pixels participate separately in the 3D model since the 3D coordinates of each pixel is retrieved using only its phase information as described in Sect. 3. Thus, we propose to create for each camera three 2D maps defined by the X, Y and Z coordinates of the 3D points. The optimization algorithm and the deblurring are applied for each camera separately to compute high-resolution images of X, Y, Z and texture from the low-resolution images.

To achieve this, we need to solve an optimization problem which jointly employs a data fidelity term $E_{data}(H)$ and a regularization energy term $E_{regular}(H)$.

$$\text{minimize } E_{data}(H) + E_{regular}(H). \tag{11}$$

$E_{data}(H)$ is defined by Eq. (12) and measures agreement of the reconstruction H with the aligned low resolution data. Here, .* denotes element-wise multiplication. W_k is a banded matrix that encodes the positions of I_k assigned for the resampling step on the high resolution target grid H. G_k is a diagonal matrix containing 0 entries for all samples from I_k which are unreliable according to the non-rigid registration result. In fact, the non-rigid registration process provides a dense correspondancy list which characterizes the spatial deviation between each point in M_t with its nearest point in the deformed model M^d_{t-1}. We employ a threshold term q to affect 0 in the matrix G_k for each couple of corresponding 3D points which have a spatial deviation greater than q.

$$E_{data}(H) = \sum_{k=1}^{N} \|W_k .* G_k .* (I_k - H)\|_2. \tag{12}$$

$E_{regular}(H)$ is a regularization energy term that guides the optimizer towards credible reconstruction H. It is defined as a sum of norms as described by Eq. (13).

$$E_{regular}(H) = \sum_{u,v} \|\nabla H_{u,v}\|_2,$$

$$\nabla H_{u,v} = \begin{pmatrix} Q_{u,v}(0,1) \\ Q_{u,v}(1,0) \\ \vdots \\ Q_{u,v}(l,m) \end{pmatrix}. \tag{13}$$

$\nabla H_{u,v}$ is a combined vector of finite difference spatial gradient approximations at different scales (l, m) at a pixel position (u, v). $Q_{u,v}(l, m)$ is a finite difference computed as shown in Eq. (14).

$$Q_{u,v}(l,m) = \frac{H_{u,v} - H_{u+l,v+m}}{\sqrt{l^2+m^2}}. \tag{14}$$

We obtain for each camera a high-resolution 3D point cloud using high-resolution data of X, Y and Z. The final high-resolution 3D point cloud is retrieved by merging the left, middle, and right obtained 3D models which are already matched since all of them contain the 3D sparse point cloud.

5 Experimental Results

The stereo system hardware is formed by three network cameras with 30 fps and a 480 × 640 pixel resolution and a LCD video projector. The computed 3D models have a resolution of approximately 30000 points. Figure 4 presents the primitives extracted and the reconstruction steps to create one facial 3D view with neutral expression from only two cameras.

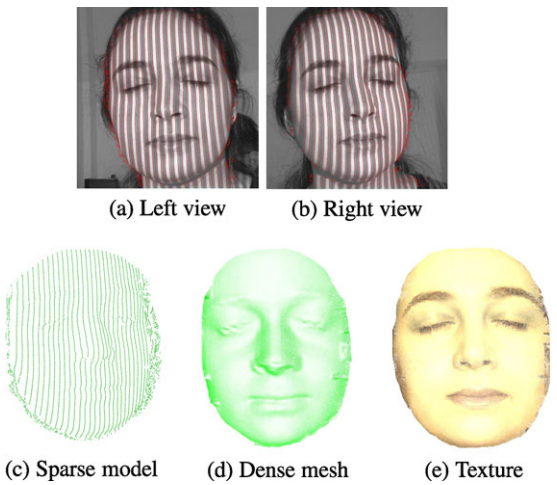


Fig. 4 Reconstruction steps to create one facial 3D view from two cameras

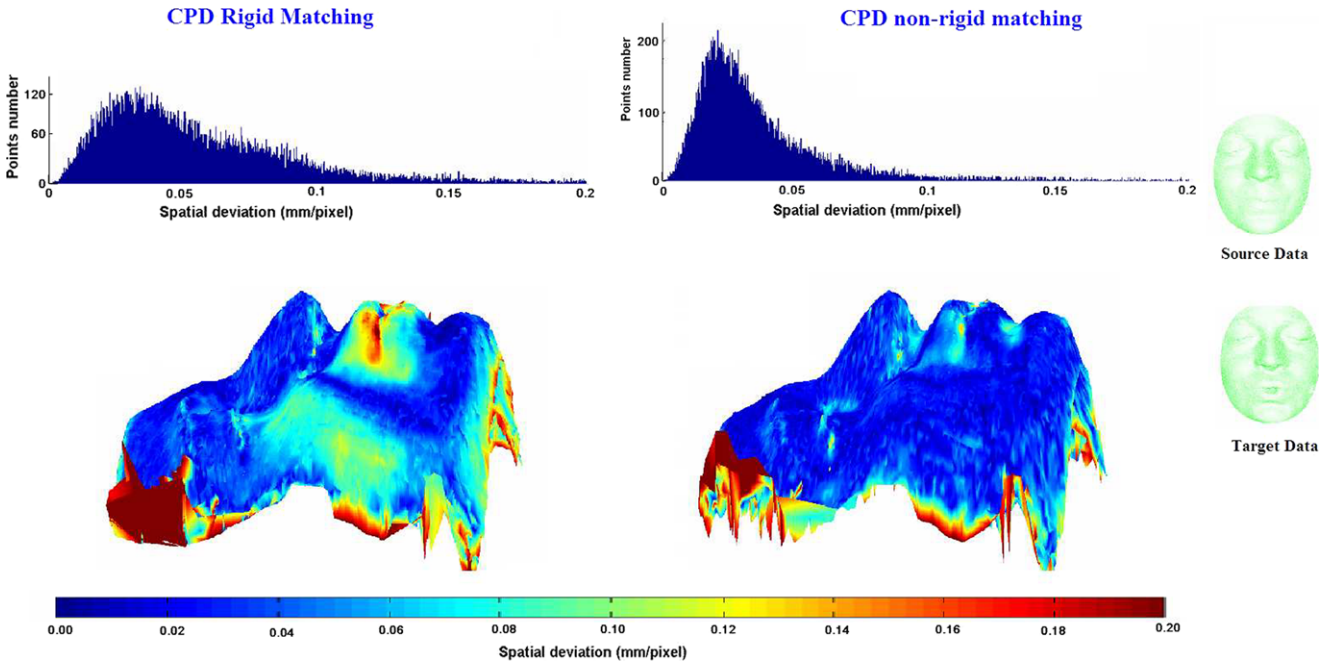


Fig. 5 The spatial deviation distributions and color cards to evaluate both Rigid and non-rigid CPD registrations

The precision of the reconstruction is estimated using a laser 3D face model scanned by a MINOLTA VI-300 non-contact 3D digitizer. We perform a point-to-surface variant of the 3D rigid registration algorithm ICP (Iterative Closest Point) between a 3D face model provided by our approach and a laser 3D model of the same face. The mean deviation obtained between them is 0.3146 mm.

5.1 Non-rigid Registration Evaluation

The proposed approach needs a non-rigid registration step between the 3D frame F_t and its preceding 3D frame F_{t-1}. The CPD performs an efficient registration in the presence of non-rigid deformations. To evaluate the CPD non-rigid registration, we propose to map a color card on the 3D frame F_t in which the color of each 3D point describes the spatial deviation separating it from its corresponding 3D point in the frame F_{t-1} after registration. Thereby, Fig. 5 illustrates the spatial deviation after rigid and non-rigid registration of F_t and F_{t-1}.

Also, Fig. 5 shows the spatial deviation distribution for both rigid and non-rigid process. The rigid registration provides a mean spatial deviation of 0.0616 mm/pixel and a standard deviation of 0.0566 mm/pixel. The non-rigid registration provides a mean spatial deviation of 0.0387 mm/pixel and a standard deviation of 0.0371 mm/pixel. Moreover, the color cards the non-rigid registration efficiency to deal with expression variation by minimizing locally and smoothly the spatial deviation in the mouth region.

We suggest to study the physical consistency of the proposed non-rigid scheme in the sense that each point of the first point cloud is the same physical point as its homologous point in the second point cloud elected by the non-rigid registration or not. To achieve this, we propose to use a standard 3D video face database mostly because our actual 3D scanning system is not designed for database acquisition. We use the BU database which consists of 101 people with 58 women and 43 men. The database is characterized by an ethnic variety: 28 Asian, 8 black, 3 Latino and 62 white subjects with ages between 18 and 45 years [19, 20].

The BU database has six types of facial expressions for each subject and a textured 3D video for each facial expression and each subject. Expressions are classified under six categories: happiness, sadness, fear, disgust, anger and surprise. Thus, the database contains 606 3D textured video in AVI format. Each 3D video has a resolution of approximately 35,000 points and its corresponding texture video has a resolution of 1040×1329 pixels per frame.

To check the validity of the physical correspondence after a non-rigid registration of two successive 3D frames F_t and F_{t-1}, we locate manually a set of landmarks on F_t. F_{t-1} is considered as a source model and F_t is considered as a destination model. Thereby, the registration process consists on deforming F_{t-1} to minimize its spatial deviation from F_t. We suggest to track the landmarks location on F_{t-1} before and after its non-rigid deformation. Figure 6 shows the landmarks locations on both the destination model F_t and the source model F_{t-1} before its deformation. Figure 7 shows the landmarks locations on the source model F_{t-1} after its

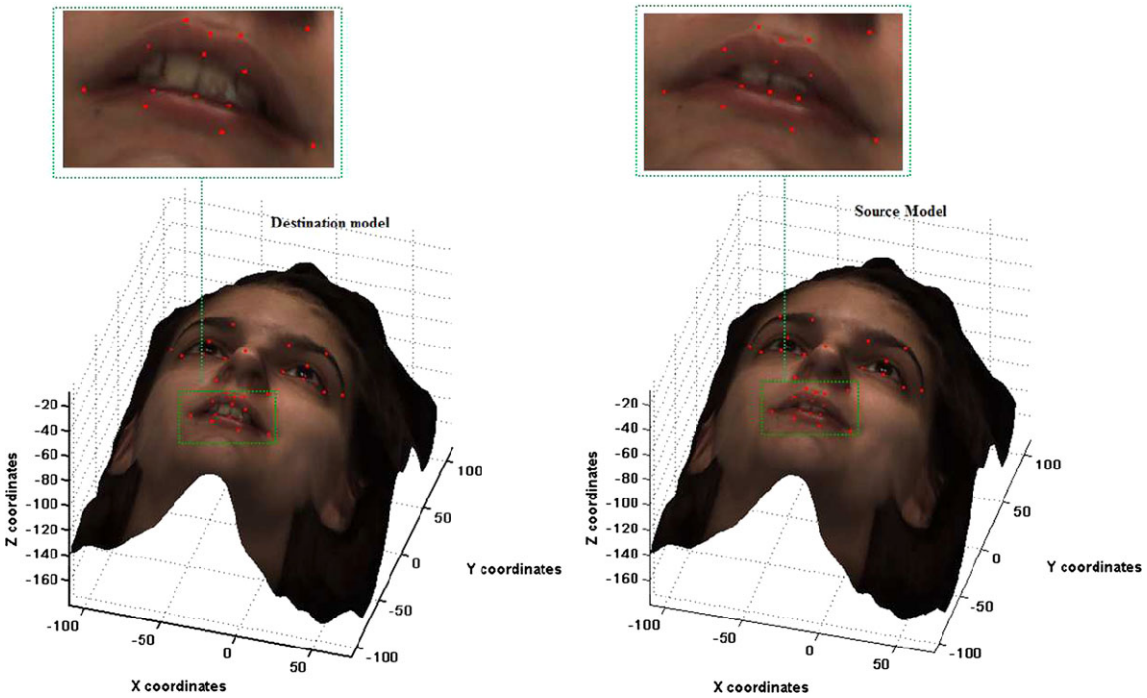


Fig. 6 Landmarks locations for BU database 3D data on both the destination model F_t and the source model F_{t-1} before its deformation

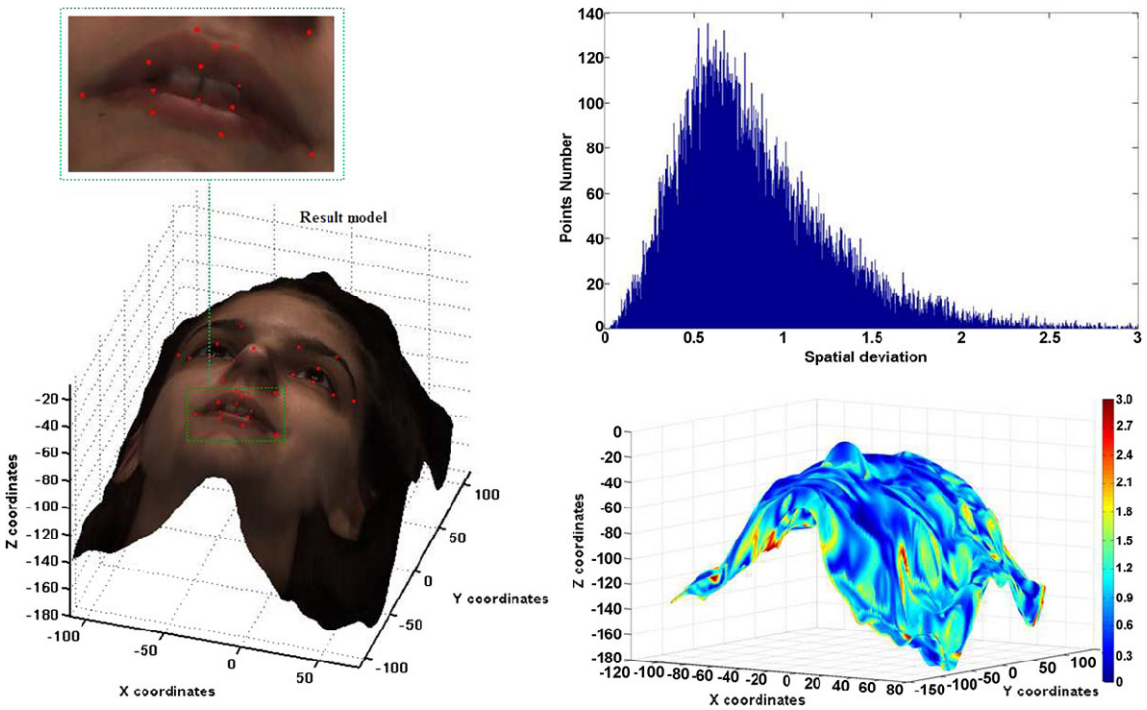


Fig. 7 Landmarks locations for BU database 3D data on the source model F_{t-1} after its deformation, the spatial deviation color card and distribution after the non-rigid registration

deformation, the spatial deviation color card and distribution after the non-rigid registration. Before the non-rigid registration, the landmarks locations on F_{t-1} are defined by their coordinates on F_t. Thereby, the landmarks are not situated on their legitimate physical locations on F_{t-1}. After the non-rigid registration, the landmarks locations on F_{t-1} are defined by the correspondancy list between F_t and the deformed F_{t-1}. Figure 7 shows that the landmarks locations

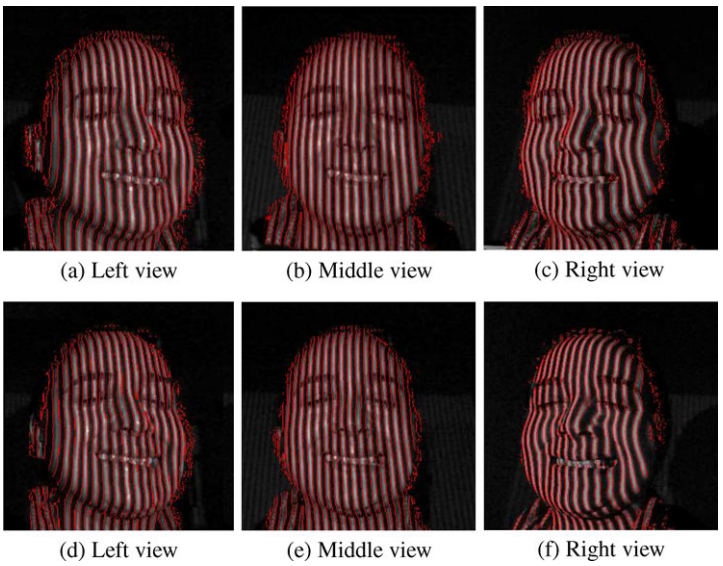


Fig. 8 Sampled primitives on *left*, *middle* and *right* views for two successive frames with an expression variation

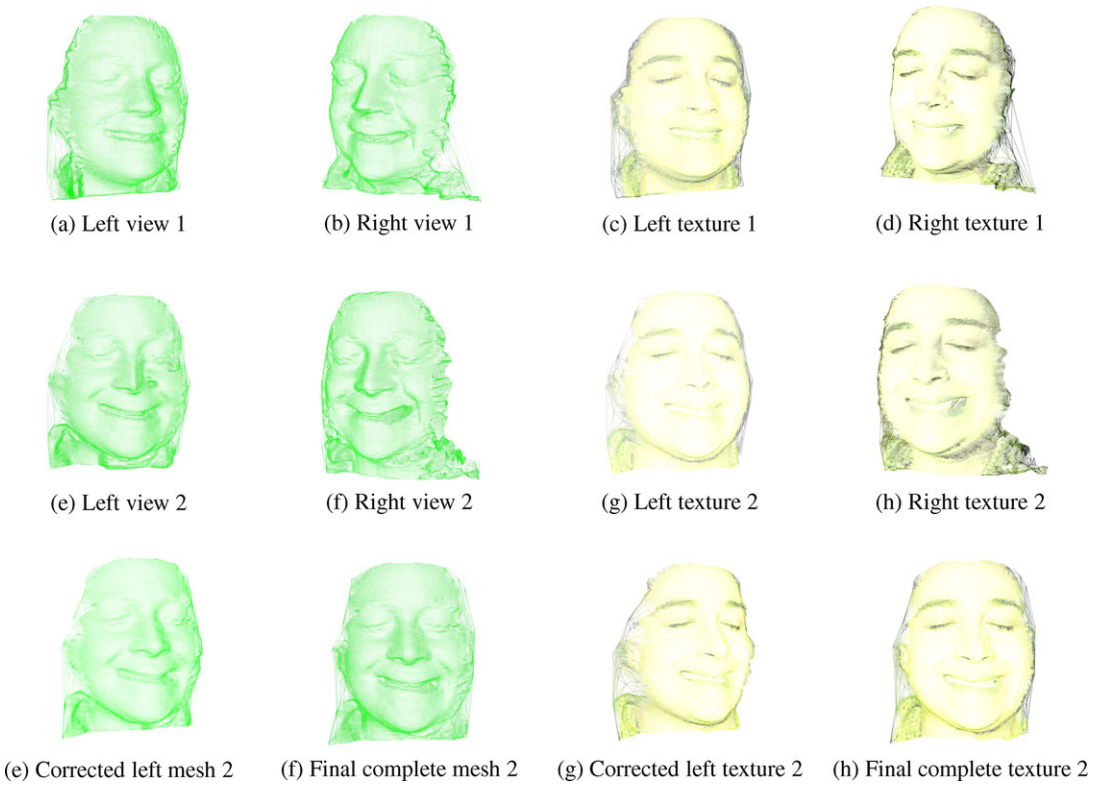


Fig. 9 3D space-time super-resolution results

on deformed F_{t-1} correspond to their legitimate physical positions.

5.2 3D Super-Resolution Results

Figure 8 presents two sets of three 2D views captured by the left, middle and right cameras. The first set is captured at time $t-1$ and the second one at time t. Each view is represented by a set of three images containing respectively the positive pattern, the π-shifted pattern and the white pattern. These 2D stereo images provide two 3D facial views. Some artifacts can be generated as shown in Fig. 9 especially for the left 3D view of the second 3D frame shown in Fig. 9.d. Here, these reconstruction errors are caused by occlusion.

To correct the 3D data, the first complete 3D frame F_{t-1} is used to correct the left and right 3D views F_t^l and F_t^r and

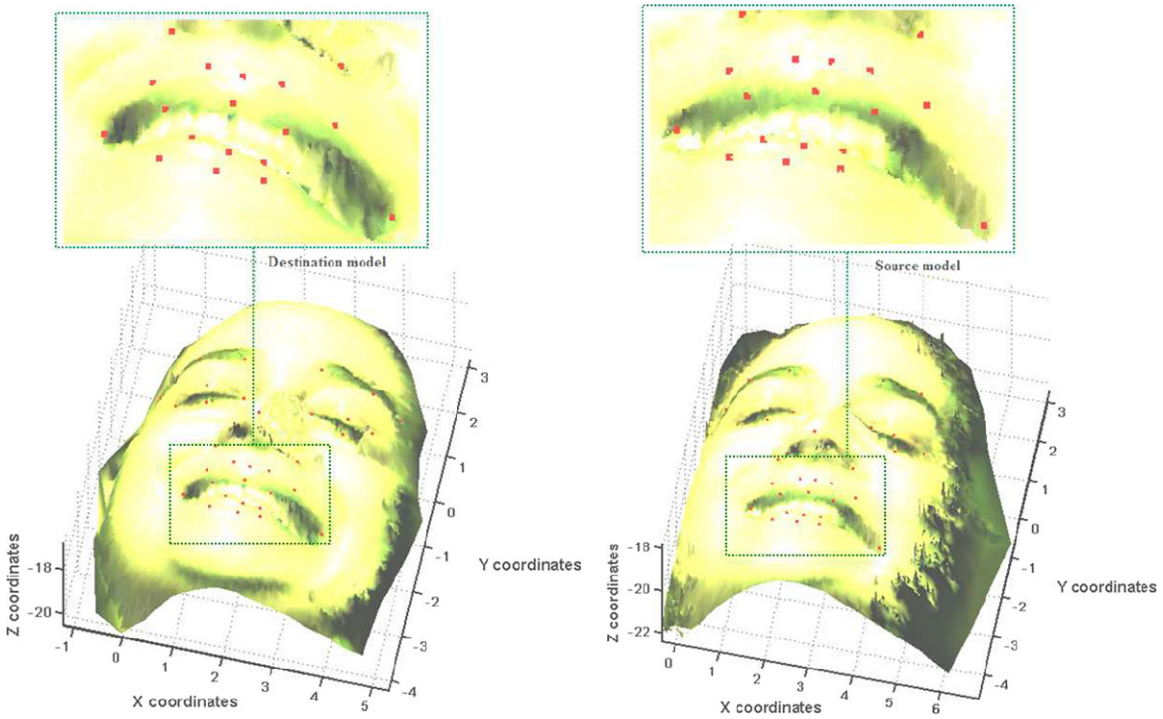


Fig. 10 Landmarks locations for our 3D data on both the destination model F_t^l and the source model F_{t-1} before its deformation

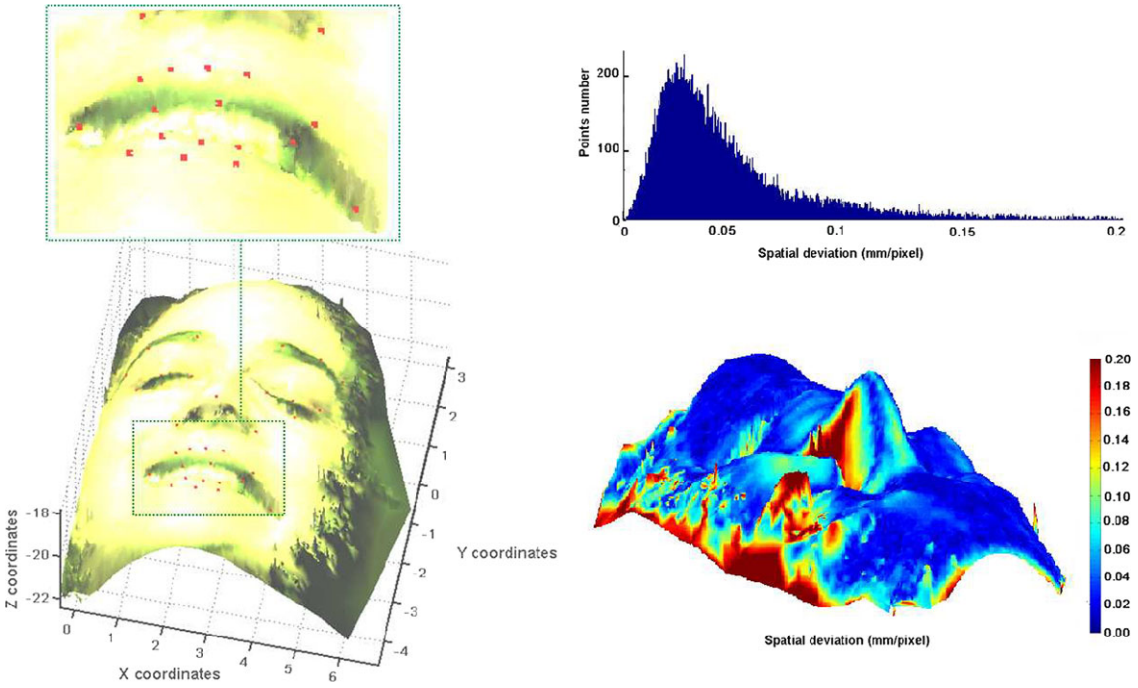


Fig. 11 Landmarks locations for our 3D data on the source model F_{t-1} after its deformation, the spatial deviation color card and distribution after the non-rigid registration

their non-rigid deformation is considered thanks to the non-rigid registration proposed in Sect. 4.1. Then, F_t^l and F_t^r are merged to make up F_t. Figure 10 shows the landmarks locations on both the destination model F_t^l and the source model F_{t-1} before its deformation. Figure 11 presents the landmarks locations on the source model F_{t-1} after its deformation, the spatial deviation color card and distribution after the non-rigid registration.

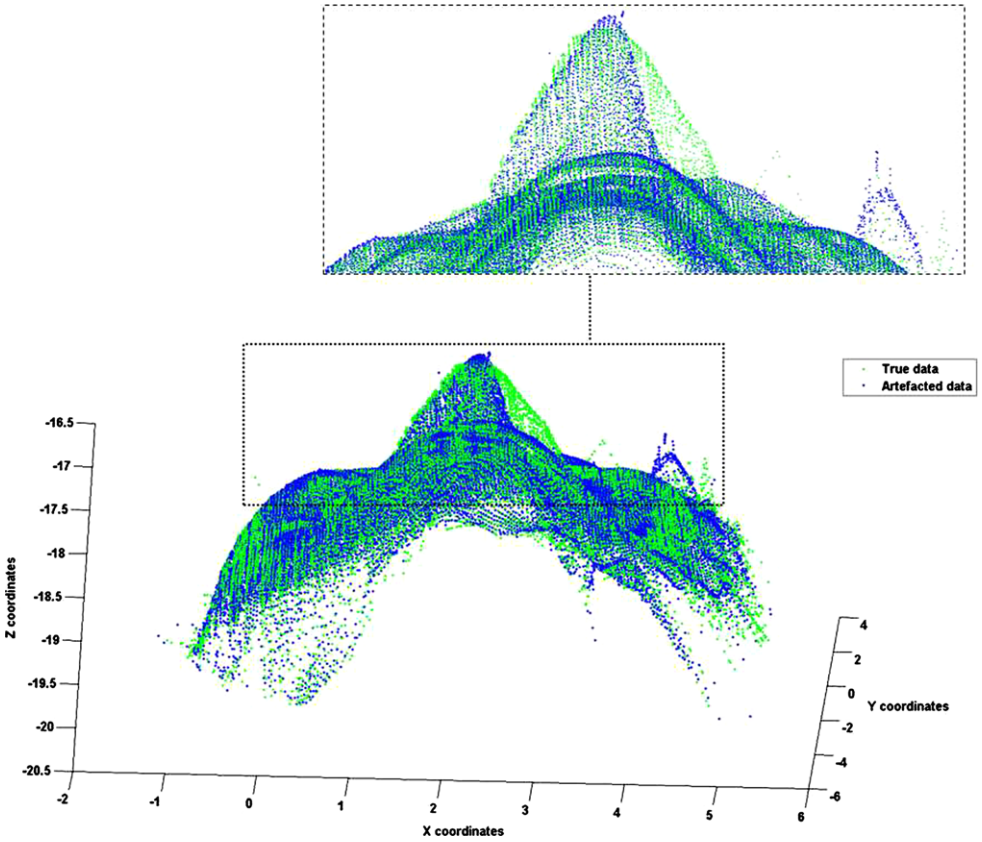


Fig. 12 Rigid registration between the true 3D model and the erroneous artifacted 3D model

The non-rigid registration algorithm CPD matchs efficiently the preceeding 3D frame F_{t-1} with the current 3D left view F_t^l with a mean deviation of 0.0493 mm/pixel. Also, the non-rigid registration localizes and clears the artifacts which represent a high spatial deviation with the preceeding 3D frame F_{t-1}.

To quantify the efficiency of the super-resolution process and to estimate the corrected error rate on the 3D data, we consider a 3D model without artifacts and introduce some artifacts and errors on it. Then, we carry on our super-resolution process to correct it. Three models are obtained: the original one, the erroneous one and the corrected one. We perform a rigid registration between the original model and the erroneous one and a rigid registration between the original model and the corrected one, as shown respectively in Figs. 12 and 13.

To generate an erroneous 3D face model, we need to simulate artifacts and noise that can appear during a real acquisition process. In our experiments, we found that areas of high curvature such as the nose are more sensitive to occlusion and can often introduce artifacts and errors generated by stereo matching. Thus, to simulate artifacts on the original model, we delete manually some left or right primitives before the stereo matching step. This disrupts the convergence of stereo matching and generates false matches and 3D artifacts or spikes. For example, deleting left primitives on the nose region simulates an occlusion and generates an artifact on the left side of the 3D scanned nose.

The spatial deviation distribution for the first rigid registration is characterized by a mean deviation of 0.0369 mm, a standard deviation of 0.0294 mm, a minimum deviation of 8.4894e-004 mm and a maximum value of 0.1997 mm. The relative mean deviation is estimated at 18.129 %. The spatial deviation between the first model and the corrected model has mean deviation of 0.0311 mm, a standard deviation of 0.0290 mm, a minimum deviation of 5.3265e-004 mm and a maximum deviation of 0.1999. The relative mean deviation is estimated at 15.332 %. Figure 14 shows some 3D frames computed using our proposed technique.

6 Comparisons and Discussion

To clarify the contributions of the proposed approach, we conducted a comparative study with similar research works from the state of the art according to the following aspects: spatial resolution, projector calibration, deformation capture, and complexity. In [9], Han et al. proposed a 3D scanning technique based on optical triangulation using stereo matching between the left and right phase images. Unlike [9], our geometrical reconstruction provides a better resolution since each pixel of each camera participates separately

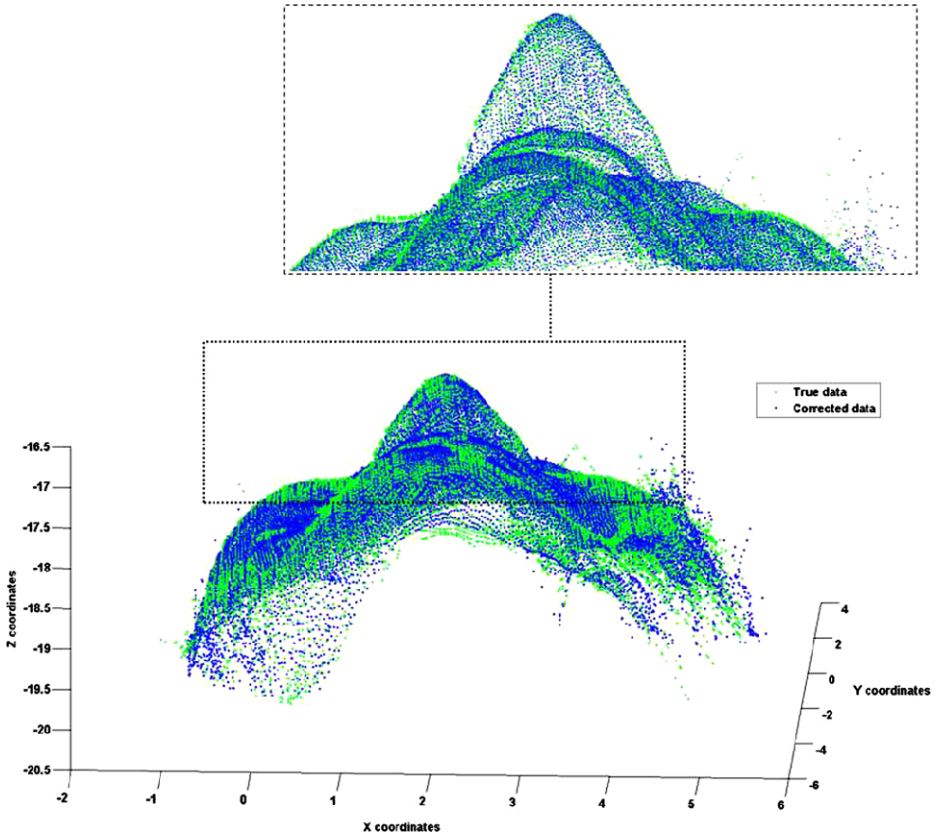


Fig. 13 Rigid registration between the true 3D model and the corrected 3D model

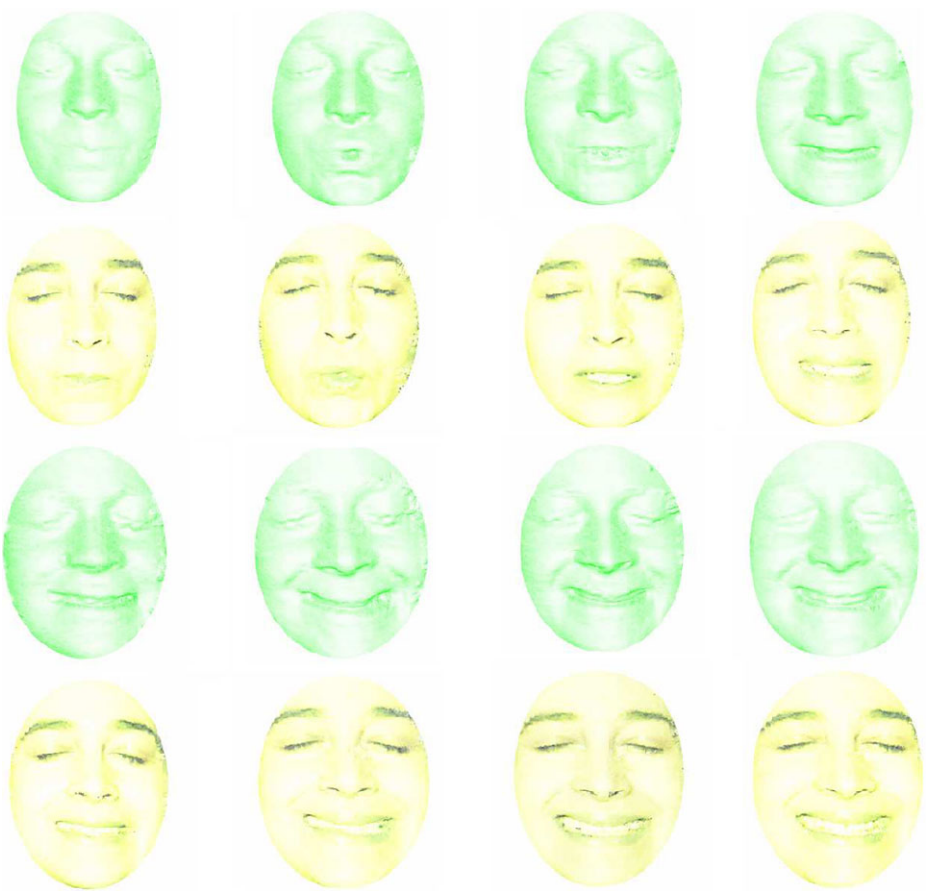


Fig. 14 Some 3D frames computed using our proposed technique

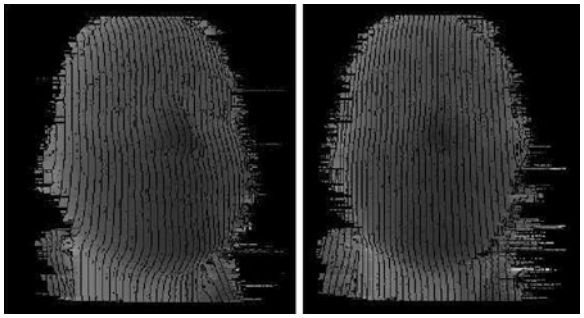


Fig. 15 The *left* and *right* disparity maps computed for two cameras 640×480

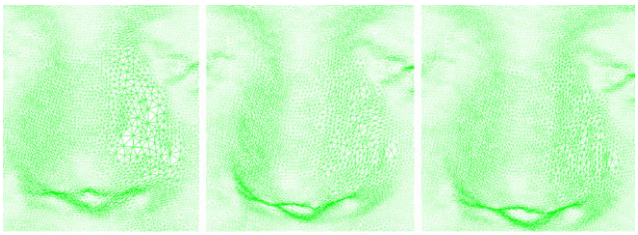


Fig. 16 3D nose scanning

in the 3D model. Two left and right disparity cards are computed as shown in Fig. 15. Figure 16.a shows the 3D point cloud computed using only the left pixels of the nose region. Figure 16.b shows the 3D point cloud computed using only the right pixels. Figure 16.c shows the fusion of the left and right point clouds for a dense 3D nose scanning.

In [18], Weise et al. introduced a stereo phase-shift technique for 3D scanning. They performed a phase unwrapping based on dense stereo matching called stereo-unwrapping and didn't precalibrate the projector. Aliaga et al. suggested a self-calibrating and multi-viewpoint framework for a photogeometric structured light technique [2, 3]. They use the initial viewing parameters estimated via a photometric method to help initialize self-calibration of a structured light system. However, their photogeometric self-calibration suffers from a sensitivity to specularities impacting the scanning precision. Also, self-calibration time ranges from an average of 15 seconds for a fast and coarse set of points to an average of 27 minutes when using all points. Unlike [2, 18], the proposed approach allows a real-time projector online localization which can be useful in certain medical and robotic applications.

Adaskevicius et al. developed a 3D scanning system based on a structured-light technique [1]. The authors designed a system composed by two calibrated cameras and a non-calibrated projector. A four step phase-shifting algorithm was chosen with $\frac{\pi}{2}$ steps. To avoid phase unwrapping, the special time-encoded binary code known as Gray code is used [15]. Illumination with a sequence of Gray code patterns yields absolute distance values, but only with poor resolution. Using four sinusoïdal and two Gray patterns matches every camera pixel to unique projected pixel. 3D shape information is obtained by optical triangulation involving left and right cameras. According to our evaluation, their system as our system, has a precision close to that of the laser scanner since it is based on structured-light (Sect. 5). However, it is less suitable for motion and deformation capture than ours as it employs more patterns.

The complexity of our stereo matching is significantly lower than the stereo-matching complexity insured in [1, 2, 9, 18] since it depends only on the number of fringes on the x-axis in contrast to [1, 2, 9, 18] where all pixels are considered for the stereo matching. Finally, it allows a sub-pixel accuracy thanks to the subsampling step while [1, 2, 9, 18] propose only a pixel resolution involving pairs of homologous pixels.

7 Conclusion and Future Work

This paper proposes a multi-camera 3D acquisition solution with a 3D space-time super-resolution scheme, which is particularly suited to 3D face scanning. It involves a pattern-based face localization approach to reduce the total processing time. This work suggests as well an online projector parametrization and does not require a camera-projector off-line calibration which constitutes a tedious and expensive task. We develop a shape+texture non-rigid registration approach to deal with the facial deformable behavior especially in the presence of an expression variation.

The proposed 3D scanning solution has its own limitations related to structured-light. Initially designed for face acquisition, various parameters such as the fringe width, the distance separating the face to the system, and the lenses of the cameras and the projector, were optimized for face and more generally for objects of similar sizes. The reuse of this technique for objects of different sizes or different distances from the system requires reconfiguration of some of these parameters. Second, our technique works only under controlled illumination. Therefore, an infrared light projection device will replace the actual visible fringe projection, which is not only intrusive and disturbing, but also leads to facial texture degradation. Finally, since we use a temporal multiplexing by projecting successive patterns onto the face, high-speed cameras should be used to improve the current performance and to scan efficiently faces or objects having a high-speed motion. We will carry out a GPU implementation of the proposed framework as well. Also, hardware synchronization between the cameras and the projector will be performed to get an efficient and real-time 3D video scanning result.

References

1. Adaskevicius, R., Vasiliauskas, A.: Three-dimensional determination of dental occlusion and facial structures using soft tissue cephalometric analysis. J. Electron. Electron. Eng. **5**(121), 93–96 (2010)
2. Aliaga, D.G., Xu, Y.: Photogeometric structured light: a self-calibrating and multi-viewpoint framework for accurate 3D modeling. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
3. Aliaga, D.G., Xu, Y.: A self-calibrating method for photogeometric acquisition of 3D objects. IEEE Trans. Pattern Anal. Mach. Intell. **32**(4), 747–754 (2010)
4. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford Univ. Press, Oxford (1995)
5. Cox, I., Hingorani, S., Rao, S.: A maximum likelihood stereo algorithm. J. Comput. Vis. Image Underst. **63**, 542–567 (1996)
6. Cui, Y., Stricker, D.: In: 3D Body Scanning with One Kinect. Conference on 3D Body Scanning Technologies, Lugano, Switzerland (2011)
7. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., Ser. B Methodol. **39**, 1–38 (1977)
8. Farsiu, S., Robinson, D., Elad, M., Milanfar, P.: Fast and robust multi-frame super-resolution. IEEE Trans. Image Process (2004)
9. Han, X., Huang, P.: Combined stereovision and phase shifting method: a new approach for 3-D shape measurement. In: Proc. of SPIE Optical Measurement Systems for Industrial Inspection VI, vol. 7389 (2009)
10. Kil, Y., Mederos, Y., Amenta, N.: Laser scanner super-resolution. In: Eurographics Symposium on Point-Based Graphics (2006)
11. Myronenko, A., Song, X.: Point set registration: coherent point drift. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 2262–2275 (2010)
12. Myronenko, A., Song, X., Carreira-Perpinan, M.A.: Non-rigid point set registration: coherent point drift. In: NIPS Conference (2007)
13. Ouji, K., Ardabilian, M., Chen, L., Ghorbel, F.: In: Pattern Analysis for an Automatic and Low-Cost 3D Face Acquisition Technique. IEEE Advanced Concepts for Intelligent Vision Systems Conference (ACIVS), Bordeaux, France (2009)
14. Rajagopalan, A., Bhavsar, A., Wallhoff, F., Rigoll, G.: Resolution Enhancement of PMD Range Maps. Lecture Notes in Computer Science, vol. 5096, pp. 304–313. Springer, Berlin (2008)
15. Salvi, J., Pagès, J., Batlle, J.: Pattern codification strategies in structured light systems. J. Pattern Recognit. **37**, 827–849 (2004)
16. Schuon, S., Theobalt, C., Davis, J., Thrun, S.: In: LidarBoost: Depth Superresolution for ToF 3D Shape Scanning. CVPR Conference (2009)
17. Tikhonov, A.N., Arsenin, V.I.: Solutions of Ill-Posed Problems. Winston and Sons, Washington (1977)
18. Weise, T., Leibe, B., Van Gool, L.: Fast 3D scanning with automatic motion compensation. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
19. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA (2006)
20. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3D dynamic facial expression database. In: IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands (2008)
21. Zhang, Z.: Flexible camera calibration by viewing a plane from unknown orientations. In: ICCV Conference (1999)
22. Zhang, S., Huang, P.S.: High-resolution, real-time three-dimensional shape measurement. J. Opt. Eng. **45**, 123601 (2006)
23. Zhang, S., Yau, S.: Absolute phase-assisted three-dimensional data registration for a dual-camera structured light system. J. Appl. Opt. **47**, 3134–3142 (2008)
24. Zhang, L., Curless, B., Seitz, S.M.: Rapid shape acquisition using color structured light and multipass dynamic programming. In: 3DPVT Conference (2002)