Appl Intell (2013) 39:761–771
DOI 10.1007/s10489-012-0410-8

# Incremental 3D reconstruction using Bayesian learning

**Ze-Huan Yuan · Tong Lu**

Published online: 23 January 2013


**Abstract** We present a novel algorithm for 3D reconstruction in this paper, converting incremental 3D reconstruction to an optimization problem by combining two feature-enhancing geometric priors and one photometric consistency constraint under the Bayesian learning framework. Our method first reconstructs an initial 3D model by selecting uniformly distributed key images using a view sphere. Then once a new image is added, we search its correlated reconstructed patches and incrementally update the result model by optimizing the geometric and photometric energy terms. The experimental results illustrate our method is effective for incremental 3D reconstruction and can be further applied for large-scale datasets or to real-time reconstruction.

**Keywords** Stereo scene analysis · Incremental reconstruction · Bayesian model · PMVS

## 1 Introduction

In computer vision, 3D reconstruction has been one of the widely researched areas in recent decades, and automatic geometric reconstruction plays a key role in automated intelligent systems [31, 34]. With the decreasing costs of video equipments, we now have the opportunity and an urgent need to run automated and accurate 3D reconstruction algorithms directly on multiple photographs or video clips. Indeed, the most important technological ingredients towards this goal are already in place. We have known that feature matching algorithms [6] can provide accurate correspondences, structure-from-motion (SFM) algorithms use these correspondences to evaluate accurate camera pose, and multi-view-stereo (MVS) methods finally reconstruct dense and accurate surface models of complex objects from a moderate number of calibrated images. Actually, the existing MVS algorithms has nearly achieved the surface coverage of about 95 % and the depth accuracy of about 0.5 mm from a set of low resolution (640 × 480) images as reported [1, 18].

MVS plays an important role in automatic acquisition of geometric objects. Existing state-of-the-art MVS algorithms can be roughly categorized into four classes: *voxel*, *mesh*, *depth maps* and *patch* based methods. *Voxel-based* MVS methods (VMVS) [2–5] represent geometry on a regularly sampled 3D grid (volume), either as a discrete occupancy function or a function encoding distance to the closest surface. Algorithms based on *deformable polygonal meshes* [7, 8] represent a surface as a set of connected planar facets and operate by iteratively evolving a surface to decrease or minimize a cost function. Approaches based on *multiple depth maps* [9, 10] model a scene as a set of depth maps and fuse individual depth maps into a single 3D model. Finally, *patch-based* MVS (PMVS) [1] algorithms output a dense collection of small oriented rectangular patches covering the observed surface obtained from pixel-level correspondences. Recently, CMVS [17] is approved effective in reconstructing from images of crowed scenes without an initialization process.

Z.-H. Yuan (✉) · T. Lu
State Key Laboratory of Software Novel Technology, Nanjing University, No. 163, Xianlin Avenue, Nanjing, Jiangsu, 210023, China
e-mail: zhyuan001@gmail.com

T. Lu
Jiangyin Institute of Information Technology of Nanjing University, No. 163, Xianlin Avenue, Nanjing, Jiangsu, 210023, China
e-mail: lutong@nju.edu.cn

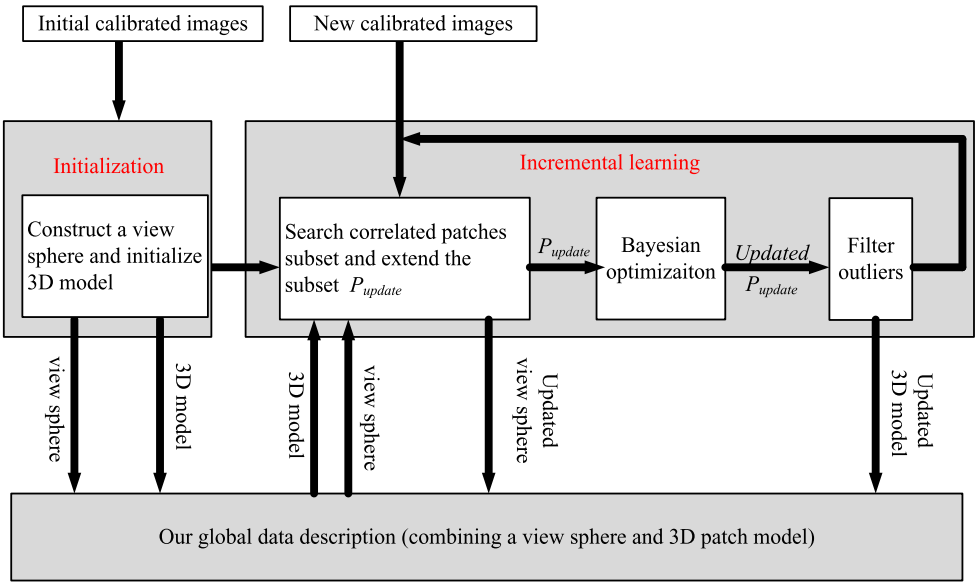


**Fig. 1** The framework of our incremental reconstruction system

However, the mentioned methods still face the following difficulties. First, they cannot handle incremental reconstruction tasks well. The input images should be well sequenced manually before reconstruction. Moreover, once a geometric object is obtained, it cannot be incrementally updated when facing a new input view image. Second, the computational cost of existing methods may rapidly increase for batch-processing of images, especially in handling huge numbers of images collected from the Internet, making it unpractical for real-time applications. Moreover, the methods also face difficulties when dealing with images of varied illuminations or scales captured by different users.

In this paper, we propose a novel algorithm aiming at incrementally reconstructing a 3D model using the Bayesian framework. We first select a group of key views uniformly distributed on our view sphere to create an initial 3D surface modeled by PMVS as stated above. Then when a new calibrated image is input, we (1) map it into a triangle on our view sphere, (2) search the correlated patches with the new input view, (3) automatically update the initial 3D model using the photometric consistency constraint and geometric smoothness priors under the Bayesian inference framework, and (4) filter patches estimated as outliers according to the visibility and photometric constraints. Note that once a new image is added, more geometric details can be extracted and integrated to incrementally optimize the final 3D model (see Fig. 1).

Our method has two main contributions. First, we propose a novel incremental 3D reconstruction framework, which makes full use of new views to incrementally update and extend an existing 3D model. As a result, the reconstruction process is more efficient and convenient, and is especially useful for automatic 3D reconstruction from a large number of real-life images or videos and real-time reconstruction. Second, to our knowledge, no previous work has attempted to reconstruct 3D dense models using the Bayesian learning framework, where pixel-level information and geometric level constraints are well integrated to optimize the final model. As a result, the reconstruction accuracy can be effectively improved.

The rest of the paper is organized as follows. Section 2 first introduces the related work. Then Sect. 3 gives details of our system. Experiments and discussions are shown in Sect. 4. Finally Sect. 5 concludes our work and the future improvements.

## 2 Related work

3D off-line reconstruction from images and video has achieved an impressive performance as mentioned in Sect. 1. Comparably, few works focus on incremental reconstruction from on-line videos or asynchronously input images, which may appear in many practical applications, such as Robot navigation, Web-based reconstruction and so on.

The existing efforts on on-line 3D reconstruction can be roughly categorized into two classes. The first class aims to compute camera pose and reconstruct spare 3D points incrementally for each frame or image. SLAM [21, 22] and SLAM based methods [20, 32, 33] are powerful techniques to achieve this target. Actually, these methods were originally designed to locate camera pose of a camera mounted on Robots moving around in an unknown scene and obtain visual odometry or sparse geometry of its surrounding environment for each frame. [19] extends it to a single uncontrolled camera in a small workspace. As a key component, the performance of SFM influences the overall effects. The development of on-line SFM systems [28, 29] for large

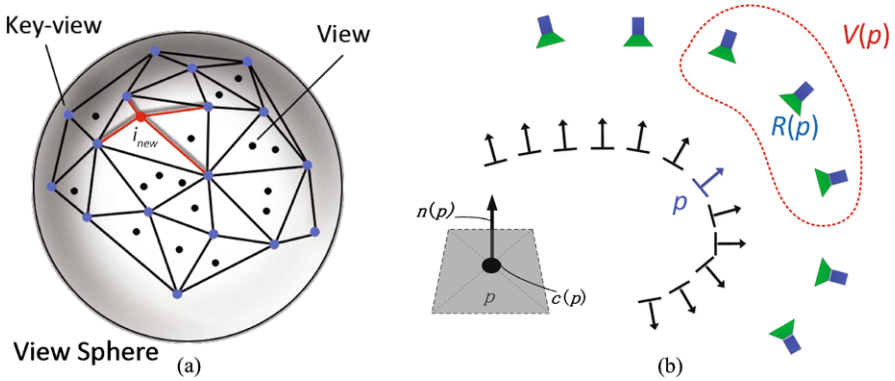


**Fig. 2** (**a**) The view sphere. (**b**) The patch model

scene reconstruction may boost the maturity of such methods; the other methods use the system of the first class as a front-end and incrementally construct a global consistent 3D model.

Recently, Merrell et al. [23] and Pollefeys et al. [24] published real-time methods using an extended plane sweeping stereo technique to reconstruct a noisy depth map for each frame and fuse these depth maps to a compact and dense map. Although having impressive results, the methods needed hardware GPU and fusing depth maps is performed at the end, making incremental updating impossible.

Other alternative approaches [25, 26] maintain a global 3D model calculated from sparse 3D feature points via Delaunay triangulation and free-space carving. When new features are added, the model is updated according to the free-space consistency. However, these methods only fuse new features to such a global model and never improve the older ones as new images and frame inputs. In addition, the simple free-space consistency, namely visibility consistency essentially, doesn't suffice to make use of the new images. Moreover, based only on features from SLAM based system and lack of extending procedure, the model is relatively sparse. Recently, another method named ProFORMA [27] was developed for on-line reconstruction. The system combines a probabilistic voting scheme and the traditional free-space constraints to raise the robustness. Nevertheless, the system also faces the same incremental problem.

## 3 Our method

In this section, we give our incremental reconstruction algorithm in details. Our method can be briefly summarized as the following four steps:

1. Map the given multi-view images set $I_{source}$ to a view sphere $S_{initial}$ and select uniformly distributed key views to initialize a 3D model;
2. For each new input image $i_{new}$, map it to $S_{initial}$ and search its related patches set $P_{update}$ on the 3D model;
3. Re-calculate the patches of $P_{update}$ using the Bayesian learning framework to incrementally refine the 3D model.
4. For all updated patches in $P_{update}$, check their visibility and photometric consistency and filter out those patches estimated as outliers.

Steps 2 to Step 4 are repeated until there are no new input images. Note that in Step 2, only a subset $P_{update}$ (named *seed patches set*) on the previous 3D model is chosen to be updated for any new input image rather than all the patches on the model. It is based on the following fact that in each incremental recursion step, the existing patches on the previous 3D model may have different *correlations* to $i_{new}$ and we need not update those patches having low *correlations*. For example, there is no (or too low) *correlation* between $i_{new}$ and another patch that is completely invisible to it. This helps reduce the computational cost, simultaneously without losing accuracy in our incremental reconstruction. The overall frame of the system can be seen in Fig. 1.

### 3.1 Initialize a 3D model

Given a calibrated image set $I_{source}$, we need first select an image subset uniformly distributed in different viewpoints to reconstruct an initial 3D model. The initial key views are selected as follows: (1) map each view image in $I_{source}$ to a view sphere $S_{initial}$ (see Fig. 2(a)), with its coordinate determined by the corresponding image plane, namely the normalized principal axis vector obtained from its projection matrix, and (2) sample key views uniformly across the sphere. Note that a point corresponding to a key view on the sphere actually represents a view image.

Next, we triangulate $S_{initial}$ by grouping the neighboring key views on it into triangles using the Delaunay Triangulation algorithm [11]. 3D patch model $\overline{S}$ can be simultaneously reconstructed using [12] from key views, reflecting an initial geometric contour of the target object. Note that the geometric contour is reconstructed using the patch-based approach [1], where a 3D surface is covered by plenty of patches, and a patch $p$ is essentially a local tangent plane approximation of the surface. A patch $p$ here has three geometric attributes (see Fig. 2(b)): $c(p)$, $n(p)$ and $R(p)$, where

$c(p)$ denotes the geometric center, $n(p)$ is the unit normal vector oriented toward the camera observing it, while a reference image $R(p)$ is an image chosen from $V(p)$ where $p$ is *truly* visible on the condition that the retinal plane of $R(p)$ is nearly parallel to $p$ within a tiny distortion.

As a result, a triangulated view sphere and a 3D patch model are obtained as the initializations of our incremental updating system.

## 3.2 Search related patches for a new input image

In our incremental reconstruction step, we first search a corresponding patch subset from the previous 3D model for any new input calibrated image, and then extend the subset to make the model more uniform and well-sampled.

### *3.2.1 Search seed patches for any input image*

To find the seed patches $P_{update}$ for any incrementally input image $i_{new}$, we first search a proper triangle $T$ on $S_{initial}$, where $i_{new}$ can be mapped into using SIFT [6] as follows:

$$T \leftarrow \arg\max_{T} \sum_{v \in T} \left| x^{v}_{i_{new}} \right| \tag{1}$$

where $x^{v}_{i_{new}}$ is a set of matches between $i_{new}$ and the key view $v$ corresponding to a vertex in triangle $T$. Then we search the correlated patch subset $P_{update}$ from the reconstructed 3D model by

$$P_{update} = \bigcup_{v \in T} \{ p \,|\, p \in \overline{S}, vis R(p) \} \tag{2}$$

Obviously, $i_{new}$ provides more useful reconstruction details for the patches in $P_{update}$ than those outside it. Then we update $S_{initial}$ as follows: (1) add a new vertex representing the new image; (2) add a pyramid of triangles by connecting the new image to three vertices of $T$, and (3) delete $T$ with $i_{new}$ located in. As a result, we can simultaneously obtain an updated view sphere (see $i_{new}$ in Fig. 2(a)).

### *3.2.2 Extend the seed patches*

Next, we extend the patch model to obtain a relatively uniform patch density along different viewpoints over the surface. The extension is associated with the orientation of the new view and the average density of the existing global surface. Note that during this process, we may create new patches under local geometric constraints to improve patch density where patches are too sparse. Our extension has the following steps:

- Estimate local density $D_p$ for every patch $p$ in 3D model. We count its neighbors $N(p)$ to evaluate the local density equivalently as follows:

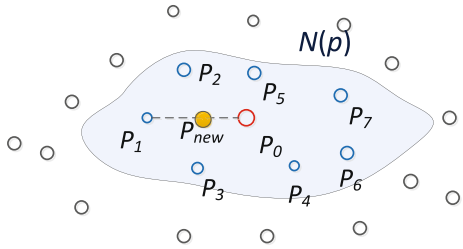


**Fig. 3** Seed patches extension, where $P_{new}$ is generated along the line combining a seed patch $P_0$ and one of its neighbors $P_1$

$$N(p) = \{ p' \,|\, p' \in \overline{S}, \left| \left( c(p) - c(p') \right) \cdot n(p) \right| + \left| \left( c(p) - c(p') \right) \cdot n(p') \right| < \rho \} \tag{3}$$

$$D_p = \left| N(p) \right| \tag{4}$$

where $\rho$ can be computed relating to the distance at the depth of the center of $c(p)$ and $c(p')$ corresponding to an image displacement of $u$ pixels in $R(p)$ ($u = 2$ in our experiment);

- Compute the global average density $D_g$ by averaging all estimated local densities;
- For every seed patch in $P_{update}$ with its local density less than $0.5^{*} D_g$, use SMOTE [13] to oversample new ones whose initialization can be seen in Table 1 between the seed patch and its neighbors (see Fig. 3). As a result, the original geometric constraints can be well maintained;
- Add the new patches into $P_{update}$.

## 3.3 Incremental surface reconstruction using Bayesian learning

This subsection introduces the Bayesian model used in our incremental reconstruction. We aim at discovering the photometric consistency and geometric smoothness constraints to obtain high-quality incremental reconstruction results.

Suppose $i_{new}$ is a *measurement* to our camera from the real scene modeled by PMVS in our method. Let $S$ be the real scene to be modeled, we need reconstruct the most likely surface $S_{MAP}$ given the *measurement* $i_{new}$. This can be achieved by maximizing the Bayesian posterior (MAP) probability $p(S|i_{new})$ in the solution space $\boldsymbol{\Omega}$

$$p(S|i_{new}) = \frac{1}{Z} p(i_{new}|S) p(S), \quad S \in \Omega \tag{5}$$

$$S_{MAP} = \arg\min \left( -\log p(i_{new}|S) - \log p(S) \right) \tag{6}$$

in order to reduce the parameter dimensions, we constraint $\boldsymbol{\Omega}$ to the expanded patches subset $P_{update}$ as mentioned in Sect. 3. Note that the constant related to $Z$ is ignored in (6). $p(i_{new}|S)$ specifies the likelihood of the *measurement* $i_{new}$ agreeing with $S$. In other words, it measures how well the normal and coordinate of a patch match the real surface according to the information hidden in $i_{new}$ and the other correlated images. It can be defined by the use of photometric

**Table 1** The incremental algorithm

Input: $S_{initial}$ and 3D patch model $\overline{S}$ reconstructed by PMVS
Output: an improved well-sample, high-resolution and more accurate patch model

While Inputting an image $i_{new}$
  Locate $i_{new}$ in $S_{initial}$ and find a corresponding triangle $T$ using SIFT
  For any $p$ in the 3D patch model
    $N_p \rightarrow \{p' | p' \in \overline{S}, |(c(p) - c(p')) \cdot n(p)| + |(c(p) - c(p')) \cdot n(p')| < \rho\}$
    $D_p \rightarrow |N(p)|$
    $P_{update} \rightarrow \bigcup_{v \in T} \{p | v \text{ is } R(p)\}$
  Update $S_{initial}$
  Compute $D_g$ by averaging all local densities
  For any $p$ in $P_{update}$
    If $D_p < 0.5 * D_g$
      Generate a new patch $k$
      $c(k), n(k) \rightarrow$ oversampling method $smote(N_p, sample\text{-}rate, p)$.
      $R(k) \rightarrow R(p)$
      $V(k) \rightarrow V(p)$
      Add $k$ into $P_{update}$
  For any patch $p$ in $P_{update}$
    $c(p), n(p) \leftarrow \arg\min(\lambda E_1 + \zeta E_2 + \eta E_p)$, $p \in P_{update}$
  Update $R(p)$ and $V(p)$ similar to [1]
  Remove outliers $\{p | p \in P_{update}, V(p) < \alpha \text{ or } E_p > \beta\}$
end while

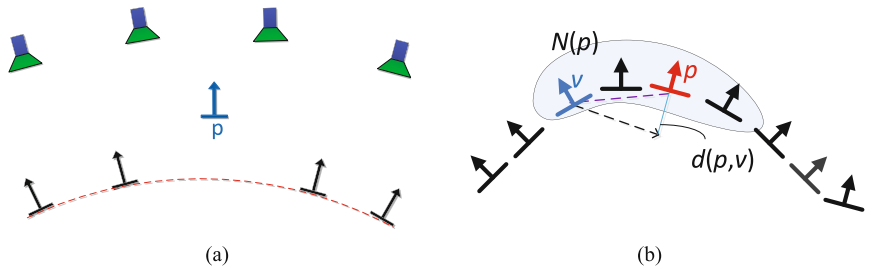


**Fig. 4** Geometric smoothness terms. (**a**) The *blue patch* $p$ is an outlier; however it has a continuous normal with its neighboring patches. (**b**) $d(p, v)$ is the absolute distance between two patches $p$ and $v$ along $n(p)$

discrepancy function [1], which we choose to express the photometric consistency:

$$p(i_{new}|S) \propto \exp(-\eta E_p) \tag{7}$$

$$Ep = \frac{1}{|S|} \sum_{p \in S} \frac{1}{|V(p)| - 1} \sum_{i \in V(p)/i_{new}} h(p, i_{new}, i) \tag{8}$$

where $\eta$ is a control coefficient, and $h(p, i_{new}, i)$ is equal to one minus the pair-wise normalized cross correlation concerning to the patch projection into images $i_{new}$ and $i$, respectively.

We use two constraints to define the prior $p(S)$:

$$p(S) \propto \exp\big(-\{\lambda E_1 + \zeta E_2\}\big) \tag{9}$$

where $E_1$ and $E_2$ are two geometric smoothness energy terms, and $\lambda$, $\zeta$ are weighted coefficients. $E_1$ is used to assure the smoothness of the reconstructed surface. For a natural 3D object, we can model its surface smoothness by accumulating sub-linear potentials of surface curvature similar to [14]. Concretely, we define $E_1$ as follows:

$$E_1 = \frac{1}{|S|} \sum_{p \in S} \frac{1}{|N(p)|} \sum_{v \in N(p)} f(p, v) \tag{10}$$

$$f(p, v) = \sqrt{\big(n(p) - n(v)\big)^T \big(n(p) - n(v)\big)} \tag{11}$$

where $N(p)$ is the neighboring patches set of $p$ defined in (3) and $f(p, v)$ is the square-root potential with $f(p, v) = 0$ if $n(p) = n(v)$ and positive otherwise.

However, there still may exist exceptions even Eq. (10) is met. For example, in Fig. 4(a), the patch $p$ is an outlier while having well sub-linear continuous relations with normals of its neighbors in $N(p)$. Considering although such a patch has a continuous normal, its geometric location is far away from the real surface, we use another geometric smoothness energy term $E_2$ to minimize such error as follows:

$$E_2 = \frac{1}{|S|}\sum_{p\in S}\frac{1}{|N(p)|}\sum_{v\in N(p)} d(p,v) \tag{12}$$

$$d(p,v) = \left|n(p)\cdot\left(c(v)-c(p)\right)\right| \tag{13}$$

where $d(p,v)$ is the distance between two patches $p$ and $v$ along $n(p)$ (see Fig. 4(b)).

This minimization problem requires us to adjust $c(p)$ and $n(p)$ for any patch $p$ in $S$ from the initial value to the final convergent solution. It is actually a sparse energy minimization optimization problem. To simplify the complexity and reduce the dimension of variables, we constrain $c(p)$ lie on a ray to assure the projection into $R(p)$ does not change. Simultaneously, we model $n(p)$ with Euler angles. Thus for every patch, only three parameters participate in the optimization problem, greatly reducing the dimension of the solution space and improve stability in the search process. We use the conjugate gradient descent to solve the global optimization problem. In this process, the derivatives for geometric smoothness prior can be directly computed and those for the photometric consistency term are currently estimated numerically.

Considering input images may be similar to each other, such as adjacent frames from the same video, we need to further group those adjacent view images and use each image group to refine the 3D model, to avoid offering redundant information and further reduce reconstruction cost. Following the intuition, we replace the measurement $i_{new}$ with an image group $G_{new}$ in the Bayesian framework and update Eq. (8) by

$$E_p = \frac{1}{|S|}\sum_{p\in S}\Bigg(\frac{1}{|V(p)|-|G_{new}|}\sum_{\substack{i\in V(p)-G_{new},\\ i_{new}\in G_{new}}} h(p,i_{new},i) + \frac{1}{(|G_{new}|)\cdot(|G_{new}|-1)/2} \times \sum_{\substack{i_{new1},i_{new2}\in G_{new},\\ i_{new1}\neq i_{new2}}} h(p,i_{new1},i_{new2})\Bigg) \tag{14}$$

Additionally, we search the Bayesian solution space $\boldsymbol{\Omega}$ or $P_{update}$ for each group. Actually, as mentioned in Sect. 3.2.1, it follows the intuition that we first search the proper triangles and the correlated patches for each new view in $G_{new}$ in the same way with Eq. (1) and Eq. (2), and then the correlated patches in $G_{new}$ are actually a union of these correlated patches, namely

$$P_{update} = \bigcup_{i_{new}\in G_{new}} P_{update\_inew} \tag{15}$$

where $P_{update\_inew}$ is the correlated patches set for each new image in $G_{new}$ obtained by Eq. (2). Note that the intersection between these correlated patches set is usually of a large number due to the adjacent views from a video.

### 3.4 Filter outliers

After refining the model, we finally remove erroneous patches inconformity to the visibility consistency and photometric consensus. Due to the reasonable initialization from the 3D model with tough geometric relation before updating, few patches become outliers caused by bad local minima after refinement. Here, we handle the photometric consistency by ignoring the patches with a low photometric score calculated by Eq. (8). As to the visibility consistency, for each patch $p$ of 3D model, we compare the number of images in $V(p)$ updated according to a depth-map test similar to [1] to a preset threshold and filter it out as an outlier. To conclude, the patches satisfying

$$\left\{p|p\in P_{update}, V(p)>\alpha, E_p<\beta\right\} \tag{16}$$

are regarded as reasonable ones and reserved in the final model. At the same time, in order to improve efficiency, we perform filtering every three rounds with new images input gradually.

As a summary, our incremental updating algorithm is shown in Table 1.

## 4 Experiments and discussions

We have implemented our incremental reconstruction algorithm on the C++ platform. The datasets [15, 16] used in our experiments are shown in Table 2, together with the number of the input images, their approximate sizes, the number of the key views we choose and the patch number of the reconstructed initial model using PMVS [12]. In our incremental processes, we set $\lambda$, $\zeta$, $\eta$, $\alpha$ and $\beta$ 0.3, 0.2, 0.7, 4 and 0.25, respectively. Figure 5 gives our incremental reconstruction results for these datasets.

In Fig. 5, Column (a) and Column (b) correspond to example 2D images and their initial result models reconstructed from key views, respectively. After gradually adding new images, the result models are incrementally updated, as shown in the rest of the three columns (c)–(e). It can be seen that the result models can be dynamically optimized and enriched with more details during these processes.

To evaluate our method quantitatively, we adopt the weighted sum of normalized cross correlations (NCC) $P_k$ to model the accuracy of a patch $k$, formulated as

$$P_k = \frac{1}{\sum_{i\in V(k)} r(k,i)}\sum_{i\in V(k)}\frac{1-h(k,R(k),i)}{r(k,i)} \tag{17}$$

$r(k,i)$ is the diameter of a sphere centered at $k$ with the projected diameter equaling one pixel in $i$. The weight related to each visible image in $V(k)$ actually reflects its contribution to the patch refinement. The measurement indicates the

**Table 2** The datasets used in our experiments

| Name | Images | Image size | Key views | Initial patches |
|---|---|---|---|---|
| Toy dinosaur | 24 | 2000 × 1500 | 15 | 27267 |
| Morpheus | 24 | 1400 × 1200 | 15 | 18433 |
| predator | 24 | 1800 × 1800 | 15 | 29620 |
| Soldier | 24 | 1300 × 1400 | 15 | 13959 |
| Human skull | 24 | 2000 × 1800 | 15 | 45223 |
| temple | 312 | 640 × 480 | 209 | 32317 |

(a) 2D images (b) the initial model (c) result 1 (d) result 2 (e) result 3

**Fig. 5** Our incremental reconstruction results. (**a**) 2D sample images, (**b**) the initial 3D model, (**c**)–(**e**) the incremental reconstruction results. From top to bottom, the datasets are *dinosaur*, *human skull cast*, *Morpheus*, *predator*, *soldier* and *temple*

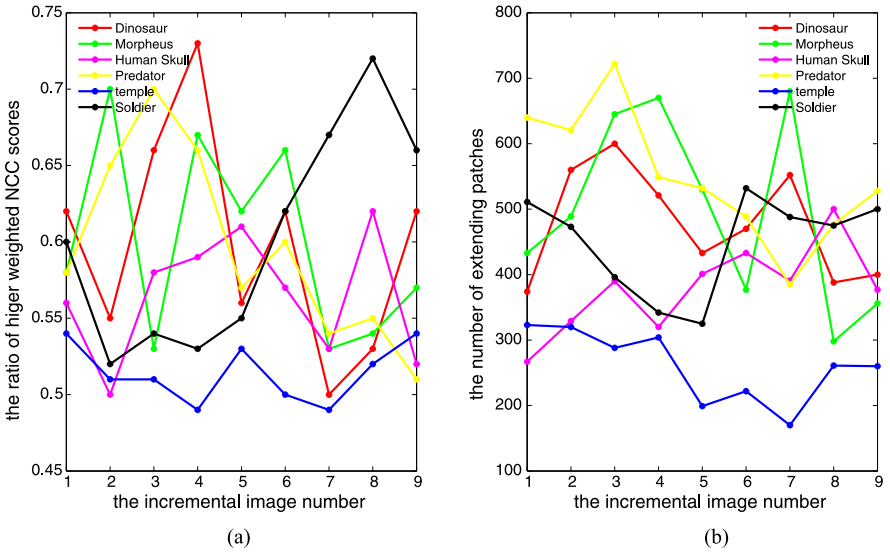


**Fig. 6** The overall statistic analysis. (**a**) The ratio of patches having higher photometric consistency scores, (**b**) the number of extended patches

**Table 3** The average accepted rate of extending patches

| Datasets | Dinosaur | Morpheus | Human skull | Predator | Temple | Soldier |
|---|---|---|---|---|---|---|
| Accepted rate | 86.9 % | 90.1 % | 91.4 % | 89.2 % | 96.0 % | 91.3 % |

**Table 4** The average accepted rate of our filter

| Datasets | Dinosaur | Morpheus | Human skull | Predator | Temple | Soldier |
|---|---|---|---|---|---|---|
| Accepted rate | 93.9 % | 95.1 % | 96.8 % | 93.4 % | 97.4 % | 94 % |

accuracy of one patch $k$ indirectly by estimating its overall photometric consistency in $V(k)$, weighing different images according to their correlations to the patch. Obviously, the larger the measurement is, the more accurate the patch is. During each incremental step, we calculate the ratios of those patches with larger weighted NCC scores in $P_{update}$ (see Fig. 6) after updating. Figure 6 is a discrete figure where different points on curves have no relations and can be replaced by tables if enough space is available. It can be seen that after adding a new image, the NCC accuracy of more than 50 % of its related patches improve averagely, illustrating the effectiveness of our method.

We also find that in Fig. 6(a), the ratio changes with image quality and position on our view sphere varying during the incremental reconstruction steps. It is due to that geometric smoothness term plays an important role in the optimization for poor-quality images, and thus the overall accuracy may be reduced simultaneously because of over-smoothing.

Figure 6(b) illustrates the number of extended patches in each incremental reconstruction step with the *sample-rate* 200 % in our experiments. Obviously, the number greatly depends on the viewpoint of a 2D image and more patches are necessary to be generated in sparse regions. Note that not all extended patches are finally added to the result model due to the global geometric constraints and the pixel-level information. Table 3 gives the accepted ratios for extending patches in our experiments, by averaging the statistical result of each new image. The performance of our filter is listed in Table 4, indicating that our system successfully removed erroneous patches. Note that few patches in our algorithm are estimated as outliers, showing that the 3D model become more accurate.

Then we use a bootstrapping-like approach to further evaluate the bias brought by the selection of key views. For each dataset, we repeat the overall algorithm 20 times and thus obtain 20 final 3D models. Note that the selected key views set is different from each other in each test round due to random sampling as mentioned in Sect. 3.1. Then we use the D2 3D matching method [30] to calculate their similarities. Finally, the average similarity $AS$ and variability $V$ are computed as follows:

$$AS = \frac{1}{190}\sum_{i=1}^{19}\sum_{j=i+1}^{20} S(i,j) \tag{18}$$

$$V = \sqrt{\frac{1}{190}\sum_{i=1}^{19}\sum_{j=i+1}^{20}\big(S(i,j) - AS\big)^2} \tag{19}$$

**Table 5** The average similarity and variability for each dataset

| Dataset | Dinosaur | Morpheus | Human skull | Predator | Soldier | Average |
|---|---|---|---|---|---|---|
| Average similarity (*AS*) | 0.878 | 0.903 | 0.926 | 0.895 | 0.914 | **0.9032** |
| Variability (*V*) | 0.071 | 0.052 | 0.008 | 0.083 | 0.074 | **0.0576** |

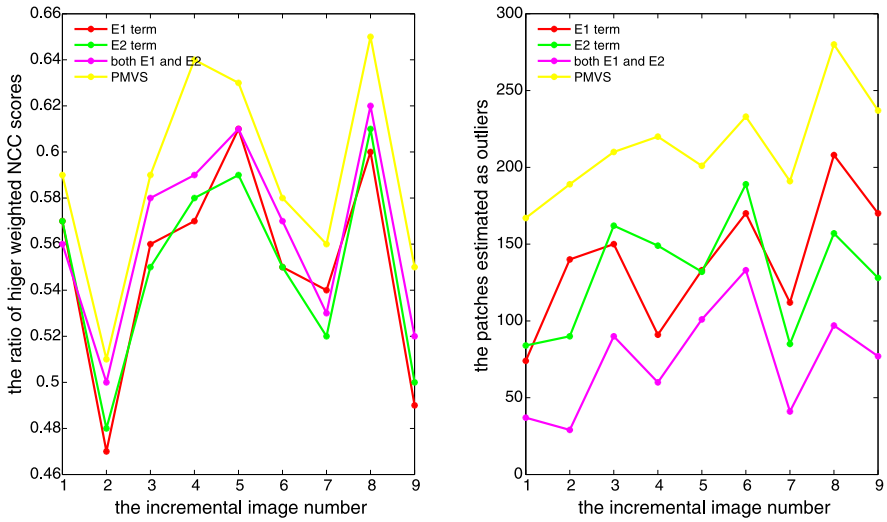


**Fig. 7** The comparison using different priors. (**a**) The ratio of patches having higher photometric consistency scores for three methods. (**b**) The patches estimated as outliers by our filter in each incremental step

where $S(i, j)$ denotes the similarity between two final models obtained in the $i$th and $j$th test round respectively. Note that a high average similarity or a low variability in general indicates the accuracy of the reconstructed 3D models or the robustness of our algorithm, respectively. The experimental results for each dataset are listed in Table 5. From the statistics, we find that either the average similarities or the variabilities for different datasets are good enough, reflecting the robustness and effectiveness of our algorithm.

We also compare our methods by adopting different smoothness combinations as: (1) a uniform prior, (2) only $E_1$, (3) only $E_2$, and (4) both $E_1$ and $E_2$, on the same dataset *human skull*. Note that the uniform prior doesn't have information and contribute to the refinement of patches. Thus the condition (1) degenerates to the traditional PMVS, which just updates patches by optimizing the photometric consistency. From the results shown in Fig. 7(a), the traditional PMVS, namely the condition (1), performs better than the other three conditions which have almost the same effects though the prior combining $E_1$ and $E_2$ seems slightly better with higher ratios of well improved patches. However, due to the lack of available constraints, it's easy for a patch to be trapped in bad local minima in the traditional PMVS, manifesting a higher measurement but incongruity to the overall geometric constrains, namely that many outliers exist in the traditional PMVS as shown in Fig. 7(b). Figure 7(b) also shows the method with the two smoothness terms together generates few outliers after updating than with either $E_1$ or $E_2$ by avoiding bad local minima effectively using enough reasonable constraints. The experiments also show that $E_1$ and $E_2$ can complement with each other and work together to guide the optimization to converge to the optimal solution.

From the comparison with the traditional PMVS algorithm, we find our method integrates the photometric consistency and geometric prior successfully using the Bayesian scheme and well improves the entire 3D model by maximizing the posterior to find a better solution. We also compare our method with different smoothness terms, and then verify our right choice for the geometric prior. However, our method still faces some shortcomings. First, despite easy to generalize to any *scenes*, the algorithm is mostly suitable to model *object* due to the use of a view sphere and limited uniform-views images or frames for initialization, unavailable to model large *scenes*. Additionally, although we group views and exploit their redundancy to accelerate the process as mentioned above for videos and similar images, the method still encounters a large challenge on time complexity because of the global optimization.

## 5 Conclusions

We propose a novel incremental reconstruction algorithm for calibrated multi-view stereo in this paper. Our method first initializes a 3D patch model using selected key views, and then when a new image is input interactively, seed

patches for which the new image provides useful reconstruction details are searched and then extended to make surface of the 3D target uniform. We finally end up the incremental learning under Bayesian framework. Experiments on 6 open datasets illustrate the effectiveness of our method. In addition, we also use a bootstrapping approach to verify the robustness of our method. Considering that the selected key views need distribute uniformly on the view sphere, our future work focuses on reconstructing crowded scene models directly from real-life videos in an online and incremental way to get rid of the limitation. Another improvement lies on evaluating different 3D reconstruction methods in a more comparable approach, especially for incremental reconstruction applications.

**Acknowledgements** The work described in this paper was supported by the Natural Science Foundation of China under Grant No. 61272218 and 61021062, the 973 Program of China under Grant No. 2010CB327903, and the Program for New Century Excellent Talents under NCET-11-0232.

## References

1. Furukawa Y, Ponce J (2010) Accurate, dense, and robust multi view stereopsis. IEEE Trans Pattern Anal Mach Intell 32(8):1362–1376
2. Pons J-P, Keriven R, Faugeras OD (2007) Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. Int J Comput Vis 72(2):179–193
3. Tran S, Davis L (2006) 3d surface reconstruction using graph cuts with surface constraints. In: European conference on computer vision (ECCV), pp 219–231
4. Vogiatzis G, Torr PH, Cipolla R (2005) Multi-view stereo via volumetric graph-cuts. In: Computer vision and pattern recognition (CVPR), pp 391–398
5. Hornung A, Kobbelt L (2006) Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding. In: Computer vision and pattern recognition (CVPR), pp 503–510
6. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
7. Hernández Esteban C, Schmitt F (2004) Silhouette and stereo fusion for 3D object modeling. Comput Vis Image Underst 96(3):367–392
8. Furukawa Y, Ponce J (2009) Carved visual hulls for image-based modeling. Int J Comput Vis 81(1):53–67
9. Goesele M, Curless B, Seitz SM (2006) Multi-view stereo revisited. In: Computer vision and pattern recognition (CVPR), pp 2402–2409
10. Strecha C, Fransens R, Gool LV (2006) Combined depth and outlier estimation in multi-view stereo. In: Computer vision and pattern recognition (CVPR), pp 2394–2401
11. http://www.cse.unsw.edu.au/~lambert/java/3d/delaunay.html
12. http://grail.cs.washington.edu/software/pmvs
13. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
14. Diebel JR, Thrun S (2006) A Bayesian method for probable surface reconstruction and decimation. ACM Trans Graph 25(1):39–59
15. http://www.cs.washington.edu/homes/furukawa/research/mview/index.html
16. http://vision.middlebury.edu/mview/data/
17. Furukawa Y, Curless B, Seitz SM, Szeliski R (2010) Towards Internet-scale multi-view stereo. In: Computer vision and pattern recognition (CVPR), pp 1434–1441
18. Furukawa Y, Ponce J (2009) Accurate camera calibration from multi-view stereo and bundle adjustment. Int J Comput Vis 84(3):257–268
19. Klein G, DW Murray (2007) Parallel tracking and mapping for small AR workspaces. In: International symposium on mixed and augmented reality (ISMAR), pp 225–234
20. Davison AJ, Reid ID, Molton ND, Stasse O (2007) MonoSLAM: real-time single camera SLAM. IEEE Trans Pattern Anal Mach Intell 29(6):1052–1067
21. Davison A (2003) Real-time simultaneous localization and mapping with a single camera. In: International conference on computer vision (ICCV), pp 1403–1410
22. Eade E, Drummond T (2006) Scalable monocular SLAM. In: Computer vision and pattern recognition (CVPR), vol 1, pp 469–476
23. Merrell P, Akbarzadeh A, Wang L, Mordohai P, Frahm J, Yang R, Nistér D, Pollefeys M (2007) Real-time visibility-based fusion of depth maps. In: International conference on computer vision (ICCV), pp 1–8
24. Pollefeys M, Nistér D, Frahm J-M, Akbarzadeh A, Mordohai P, Clipp B, Engels C, Gallup D, Kim SJ, Merrell P, Salmi C, Sinha SN (2008) Detailed real-time urban 3D reconstruction from video. Int J Comput Vis 78(o):2–3. 143–167
25. Lovi D, Birkbeck N, Cobzas D, Jägersand M (2010) Incremental free-space carving for real-time 3D reconstruction. In: Fifth international symposium on 3D data processing visualization and transmission (3DPVT)
26. Hilton A (2005) Scene modeling from sparse 3D data. Image Vis Comput 23(10):900–920
27. Pan Q, Reitmayr G, Drummond T (2009) ProFORMA: probabilistic feature-based on-line rapid model acquisition. In: British machine vision conference (BMVC)
28. Nistér D, Naroditsky O, Bergen JR (2004) Visual odometry. In: Computer vision and pattern recognition (CVPR), pp 652–659
29. Agarwal S, Snavely N, Seitz SM, Szeliski R (2010) Bundle adjustment in the large. In: European conference on computer vision (ECCV), pp 29–42
30. Osada R, Funkhouser TA, Chazelle B, Dobkin DP (2002) Ops shape distributions. ACM Trans Graph 21(4):807–832
31. Thang ND, Kim T-S, Lee Y-K, Lee S (2011) Estimation of 3-D human body posture via co-registration of 3-D human model and sequential stereo information. Appl Intell 35(2):163–177
32. Kang J-G, Kim S, An S-Y, Oh S-Y (2012) A new approach to simultaneous localization and map building with implicit model learning using neuro evolutionary optimization. Appl Intell 36(1):242–269
33. Bonev B, Cazorla M, Martín F, Matellán V (2012) Portable autonomous walk calibration for 4-legged robots. Appl Intell 36(1):136–147
34. Bayrak AG, Polat F (2012) Formation preserving path finding in 3-D terrains. Appl Intell 36(2):348–368

**Ze-Huan Yuan** received the BS degree in computer science from Nanjing University (2012), China, where he is currently working towards his PhD. His research interests include object discovery, scene understanding, 3D reconstruction, and multimedia. He was honored as the 2012 Excellent Undergraduate Students of CCF (China Computer Federation), China, and the best paper award from IEA/AIE (2012). He is a student member of CCF.

**Tong Lu** received the Ph.D. degree in computer science from Nanjing University in 2005. He received his M.Sc. and B.Sc. degree from the same university in 2002 and 1993, respectively. He served as Associate Professor and Assistant Professor in the Department of Computer Science and Technology at Nanjing University from 2007 and 2005, and has served as Visiting Scholar at Department of Computer Science and Engineering, Hong Kong University of Science and Technology. He is also a member of State Key Laboratory of Novel Software Technology in China. He has published over 50 papers and authored 2 books in his area of interest, and issued more than 20 international or Chinese invention patents. His current interests are in the areas of document analysis and recognition, computer vision and pattern recognition algorithms/systems.

Dr. Tong Lu was a member of ACM, IAPR, ISAI and a senior member of China Computer Federation (CCF). He is the Youth Associate Editor of Journal on Frontiers of Computer Science (FCS), and has served as the Secretary-general of CAD&CG Committee of Jiangsu Computer Federation in China since 2008. He has been member of the program committee or session chair of more than 10 international scientific conferences, and the Chair of Organization Committee of Youth Scholar Forum of State Key Laboratory for Novel Software Technology since 2010.