A Unified Framework for Performance Evaluation of 3-D Reconstruction Techniques

Ahmed Eid and Aly Farag
Computer Vision and Image Processing (CVIP) Laboratory*
University of Louisville, Louisville, KY 40292
{eid,farag}@cvip.uofl.edu
http://www.cvip.uofl.edu

Abstract

In this paper we consider the problem of performance evaluation of 3-D reconstruction techniques from calibrated sequence of images. We propose a unified framework for the evaluation process assuming the existence of ground truth data collected by a 3-D laser scanner. Firstly, we propose a new pre-evaluation technique for 3-D data registration based on silhouettes alignment. The performance of this technique is independent of the output quality of the 3-D reconstruction technique under test. This is a major advantage of our technique since the quality issue, which is not guaranteed under the evaluation topic, is a challenge for any 3-D registration technique. Secondly, we propose a geometrically based evaluation methodology. The method is based on the geometrical matching of uniformly distributed patches of both ground truth and under-test data sets. The method is insensitive to the relative density of data sets. In addition, it provides a method for localizing errors which is a crucial step for diagnosis and data fusion post-evaluation techniques. Finally, we provide experimental results on the evaluation of stereo and space carving techniques.

1 Introduction

Although the performance evaluation of 3-D reconstruction techniques is an important topic, it is not usually treated in a stand alone research or even as a main topic. This makes many of the evaluation methodologies “algorithmic” in nature, i.e. they are not independent of the algorithms under test. This, in turn, has led to the presence of unpopular and limited-use evaluation approaches. Certainly, this situation does not serve the goal of having standard, on-shelf methodologies that are able to quantify the performance of existing and future 3-D reconstruction techniques. Here, we try to formulate the problem and provide possible solutions under unified framework.

Formally, we want to solve the following problem: Given (i) a set of 3-D data points $M$ of an object generated by a 3-D reconstruction technique $X$ and (ii) a set of 3-D data points $G$ of the same object generated by a 3-D scanning device, quantify the performance of the technique $X$. In general, to solve this problem three main components should be available: (i) an experimental testbed for collecting data, (ii) pre-evaluation techniques for preparing data for the evaluation process with minimal undesirable effects on the given data, and (iii) a performance evaluation methodology. Having these components within a unified framework could ease the solution and avoid unnecessary complexities if they were treated separately.

Szeliski and Zabih [1, 2] presented a leading work under the evaluation topic by providing new metrics and methodologies for the performance evaluation of stereo techniques. In addition, Szeliski [1] presented the concept of using the images as ground truth in the evaluation of stereo and motion techniques. In our previous work, [3] we used this concept and extend the evaluation domain to include voxel-based approaches [4, 5] in addition to stereo approaches.

Scharstein and Szeliski [6] have provided a distinctive comparative survey of stereo techniques. In addition, they have conducted a serious attempt to create ground truth data sets for the stereo evaluation [7].

Another important work by [8] presented a design of an experimental setup for performance evaluation of stereo techniques in tele-presence. This setup uses a 3-D scanner for providing the ground truth for performance evaluation of stereo techniques.

Similar to [8], we have developed a vision platform that fuses commercially available optical and laser sensors for vision-based applications [9]. In contrast to [8], our system uses moving cameras not fixed cameras, hence the flexibility in collecting data. In addition, our system, in contrast to the previously mentioned research, is not of limited use to the stereo approaches.

*This work is supported by US Army under grant MABT60-02-P-0063 and Air Force Office of Scientific Research (AFOSR) grant F49620-01-1-0367.
Here, we integrate this system with per-evaluation and evaluation methodologies to quantify the performance of the 3-D reconstruction techniques under a unified framework. Firstly, we propose a 3-D registration technique to align the measured and the ground truth data sets as a pre-evaluation step. The approach does not rely on the presence of the 3-D reconstruction under test during the registration phase. This gives an advantage to our approach since the 3-D reconstruction could be of low quality that might add difficulties to any 3-D registration technique. In addition, if the actual 3-D reconstruction under test were used in the registration phase, then some errors that the evaluation process tries to investigate might disappear during the minimization step used by any 3-D registration technique.

We consider the 3-D registration step the core of any evaluation work that employees 3-D ground truth data provided that it does not affect the accuracy of the evaluation process itself. Employing an efficient 3-D registration technique could make the design of an evaluation methodology a straightforward task. In addition it helps, along with a suitable evaluation methodology, in localizing the errors in the 3-D reconstruction under test. This localization step is highly demanded by diagnosis and data integration post-evaluation techniques.

As in conventional 3-D registration techniques [10], we search for rotation and translation parameters that bring the 3-D measured data and the ground truth data into match. In the proposed approach, we equivalently search for the parameters that rotate and translate a given ground truth data such that a given set of input silhouettes is aligned with a set of ground truth silhouettes. These later silhouettes are generated by projecting the ground truth data by a virtual camera having the same calibration parameters as the actual camera. Our proposed methodology has a common feature with the work presented in [11] of using silhouettes for data alignment. The work in [11] uses silhouettes for texture mapping. The method is very approximate since it uses only one silhouette, which is not enough in most cases, to evaluate the registration parameters at a single view. This process is repeated independently at each view. To refine such solution, an additional alignment step using textures is used. However, in our work we provide an accurate solution using a set of silhouettes with no need to the refinement step of using textures.

Secondly, we propose a general testing methodology to any 3-D reconstruction from sequence of images. The methodology is based on the presence of ground truth data and assumes full alignment of the ground truth data and the measured data. The proposed methodology manipulates data as a collection of patches and assigns a quality index to each patch by extracting geometrical features of both ground truth patches and the measured patches. This gives the proposed methodology the ability to localize errors, hence permitting the application of post-evaluation diagnosis and fusion techniques.

The rest of this paper is organized as follows. Section 2 provides a brief description of the used testbed and describes the proposed registration and evaluation methodology. Section 3 provides experimental results of applying the proposed methodology to two common 3-D reconstruction techniques; stereo and space carving [4]. Section 4 provides the conclusions and the future work.

2 Performance Evaluation Framework

2.1 Data Acquisition

In our research laboratory, we have developed a general purpose platform for vision applications [9]. The setup consists of a 3-D laser scanner and a CCD camera mounted on a metal arm of multiple joints which is attached to the scanner head. A mono-color, usually black, screen is attached to the scanner head facing the CCD camera such that the screen appears as a fixed background to the object under reconstruction. The structure of the mono-color screen as well as the motion mechanism of the scanner facilitate the object segmentation task. The shaft over which the scanner head is mounted, is controlled in terms of speed and angle of rotation to capture a sequence of images $I_0, I_2, ... I_{N-1}$ at specific locations on a circular path. Applying the 3-D reconstruction technique $X$ to the acquired sequence of images, we get the measured data set $M$. The 3-D scanner is used to generate the ground truth data set $G$.

2.2 Pre-evaluation: 3-D Data Registration Through Silhouettes (RTS)

Since our goal is to evaluate the quality of a given data set $M$ of measured points generated by a given 3-D reconstruction technique $X$, the ground truth data set $G$ is required to be aligned with $M$.

Overview of the Method

Specifically, 3-D registration is of our concern since we are manipulating 3-D data. Here we present a novel technique for 3-D data registration. This technique is dedicated to the evaluation problem where we assume the availability of ground truth data. However, in some cases the measured data could be corrupted or of unknown quality which in turn complicates the registration process. In this situation, the conventional 3-D registration techniques might not succeed to solve the problem unless human intervention is assumed.
As a result another technique is needed to solve the problem. We propose a 3-D registration technique that uses silhouettes, we call it Registration Through Silhouettes (RTS).

Since we evaluate 3-D reconstruction $\mathcal{M}$ from calibrated sequence of images, a set $\mathcal{S}_{in}$ of silhouettes can be generated. In addition, we use $\mathcal{G}$ to generate another set $\mathcal{S}_G$ of silhouettes at same views as set $\mathcal{S}_{in}$ In the ideal case when $\mathcal{M}$ and $\mathcal{G}$ are initially registered, $\mathcal{S}_{in}$ and $\mathcal{S}_G$ are aligned. However, in most cases a certain transformation $\mathcal{T}$ is needed to align $\mathcal{G}$ with $\mathcal{M}$. Applying $\mathcal{T}$ iteratively to $\mathcal{G}$ to get $\mathcal{S}_G$ such that the error between $\mathcal{S}_{in}$ and $\mathcal{S}_G$ is minimal will lead to getting the best $\mathcal{T}$ that brings $\mathcal{G}$ and $\mathcal{M}$ into match.

**Error Criterion**

As a formal $\mathcal{T}(R, t)$ problem, the goal is to find the transformation $\mathcal{T}(R, t)$ where $\mathcal{R}$ is a $3 \times 3$ rotation matrix with 3 degrees of freedom (DOF): $\theta_X$, $\theta_Y$, and $\theta_Z$ and $\mathcal{t}$ is a 3-D translation vector that has 3 DOF: $t_X$, $t_Y$, and $t_Z$ such that the energy $E$ is minimal:

$$E = \sum d^2(m_i \in \mathcal{M}, T(R, t)(g_i \in \mathcal{G}))$$

where $d$ denotes the distance. Since $\mathcal{M}$ is not an ideal reconstruction and the minimization could be difficult to be performed directly in the 3-D coordinates, we reduced the problem into 2-D minimization through silhouettes.

We assume that $\mathcal{M}$ is not available at the registration phase but its calibrated silhouette set $\mathcal{S}_{in}$ is. We generate $\mathcal{S}_G$ by projecting $\mathcal{G}$ to the same views as of $\mathcal{S}_{in}$.

**The Registration Procedure**

For each iteration $i = 1, \ldots, N_{max}$, where $N_{max}$ is the maximum number of iterations, the the registration parameters ($\theta^1_X, \theta^1_Y, \theta^1_Z, t^1_X, t^1_Y, t^1_Z$), are used to find the transformed set $\mathcal{G}_{i+1}$, where:

$$\mathcal{G}_{i+1} = T(R_i, t_i)\mathcal{G}_i$$

In general,

$$T(R, t) = \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix}$$

For each set $\mathcal{G}_i$, a corresponding set of silhouettes $\mathcal{S}_G_i$ of cardinality $N_i$ is generated where:

$$\mathcal{S}_{G_i} = \{ s^l_{G_i} : s^l_{G_i} \subset \mathcal{S}_G_i, l = 1, \ldots, N_i \}$$

for each point $I^l_{G_i}(x^l_{G_i}, y^l_{G_i}) \in s^l_{G_i}$ and $g^l_i = (X^l_{G_i}, Y^l_{G_i}, Z^l_{G_i}) \in \mathcal{G}_i$ which is visible at view $l$, the following relation holds at proper projection matrix $\mathcal{P}_l$ such that:

$$\mathcal{P}_l \begin{bmatrix} x^l_{G_i} \\ y^l_{G_i} \\ 1 \end{bmatrix} = \begin{bmatrix} X^l_{G_i} \\ Y^l_{G_i} \\ Z^l_{G_i} \\ 1 \end{bmatrix}$$

where $c$ is a scalar value and

$$I^l_{G_i}(k_1, k_2) = \begin{cases} L_1, & \text{if } k_1 = x^l_{G_i} \text{ and } k_2 = y^l_{G_i}; \\ L_2, & \text{otherwise}. \end{cases}$$

where, $L_1$, and $L_2$ are two gray levels, $1 \leq k_1 \leq N_h$, $1 \leq k_2 \leq N_w$, and $N_h \times N_w$ is the cardinality of $s^l_{G_i}$.

For a sequence of $N_s$ input images $I^l$, a corresponding set of silhouettes $\mathcal{S}_{in}$ can be extracted as:

$$\mathcal{S}_{in} = \{ s^l_{in} : s^l_{in} \subset \mathcal{S}_{in}, l = 1, \ldots, N_s \}$$

such that for each point $I^l_{in}(k_1, k_2) \in s^l_{in}$

$$I^l_{in}(k_1, k_2) = \begin{cases} L_1, & \text{if } I^l(k_1, k_2) \text{ is a silhouette point;} \\ L_2, & \text{otherwise}. \end{cases}$$

The error criterion $\mathcal{E}_i$ is defined as:

$$\mathcal{E}_i = \frac{1}{N_sN_hN_w} \sum_{l=1}^{N_s} \sum_{k_1=1}^{N_h} \sum_{k_2=1}^{N_w} [I^l_{in}(k_1, k_2) - I^l_{G_i}(k_1, k_2)]^2$$

then an optimization algorithm is needed to find the solution of

$$\min_{R, t} \mathcal{E}_i \rightarrow \min_{R, t} \mathcal{E}$$

a minimization procedure is described in the next section.

**A Two-step Minimization**

We use a genetic search algorithm to minimize (9). To apply GA to our registration problem, we encoded the transformation parameters as genes, with each parameter is encoded by 16 bits. The genes are formed by concatenating six binary coded parameters, the angles of rotation; $\theta_X$, $\theta_Y$, and $\theta_Z$ and the translation components $t_X$, $t_Y$, and $t_Z$. The crossover operation occurs at multi points along the gene with probability $p_c = 0.95$. A mutation rate of 0.001 is used. Since GA maximizes an objective function, we used the following objective function $\mathcal{F}$ to be maximized:

$$\mathcal{F} = \frac{1}{\mathcal{E}}$$

Since GA is a global search method that converges at infinity, we used it only to get a primary solution to a local search method. Here we used the Nelder-Mead (NM) simplex as a local search method. It is important to note that other suitable optimization techniques can replace GAs+simplex solution such as simulated annealing etc. without affecting the validity of the RTS technique.

In the presented algorithm, the convergence of the optimization techniques and hence the performance of RTS technique depends on the selection of $\mathcal{S}_{in}$ and to what extent the silhouettes are distinct. Symmetric objects that have
similar silhouettes are of no interest under the evaluation topic. Simply, if needed they can be synthetically generated. In practice, a subset of \( S_m \) that consists of 4 orthogonal silhouettes provides enough constraints on the shape of moderate complexity objects.

### 2.3 Performance Evaluation Methodology

Since \( M \) and \( G' = T(R, t)G \) has been aligned to each other, a performance evaluation methodology can be applied to both sets to measure the similarity between them. Let \( U \) be a superset of \( M \cup G' \) that is upper bounded by the 3-D vector

\[
u_b = \max \left( \max_i(n_i \in M), \max_i(g_i' \in G') \right)
\]

and lower bounded by the 3-D vector

\[
u_b = \min \left( \min_i(n_i \in M), \min_i(g_i' \in G') \right)
\]

Assume that \( X_m \) is a set of uniformly distributed 3-D points \( x_{m, j} = 1, 2, \ldots, N_m, \) in the space bounded by \( \nu_b \) and \( \nu_b \) and \( N_m \) is a user defined parameter.

Assume that \( M \) can be expressed as:

\[
M = \bigcup_{j=1}^{N_m} M_j, \ j = 1, 2, \ldots, N_m
\]

where \( M_j \) is defined as:

\[
M_j = \{ m \in M : x_{m, j} - \Delta X \leq m \leq x_{m, j} + \Delta X \}
\]

where \( \Delta X = (\Delta x, \Delta y, \Delta z) \), and \( \Delta x, \Delta y, \) and \( \Delta z \) are elementary distances in the space whose values are determined by \( N_m, \nu_b, \) and \( \nu_b \) as:

\[
\Delta X = \frac{1}{2 \sqrt{2} N_m} (\nu_b - \nu_b)
\]

Similar definitions of \( G' \) and \( G^{j'} \) are as follows:

\[
G' = \bigcup_j G^{j'}, \ j = 1, 2, \ldots, N_m
\]

and

\[
G^{j'} = \{ g' : g' \in G', x_{m, j} - \Delta X \leq g' \leq x_{m, j} + \Delta X \}
\]

For each data subset pair \( (M_j, G^{j'}) \), we define a quality index \( Q^j \). First we compute the centroid of each subset assuming each 3-D point has unity mass:

\[
C_{M_j} = \frac{1}{\text{card}(M_j)} \sum_{i=1}^{\text{card}(M_j)} (m_i \in M_j)
\]

and

\[
C_{G^{j'}} = \frac{1}{\text{card}(G^{j'})} \sum_{i=1}^{\text{card}(G^{j'})} (g_i' \in G^{j'})
\]

where, \( \text{card} \) denotes the cardinality, then we compute the deviation of each subset as:

\[
D_{M_j} = \sqrt{\frac{1}{\text{card}(M_j)} \sum_{i=1}^{\text{card}(M_j)} d^2(m_i \in M_j, C_{M_j})}
\]

and

\[
D_{G^{j'}} = \sqrt{\frac{1}{\text{card}(G^{j'})} \sum_{i=1}^{\text{card}(G^{j'})} d^2(g_i' \in G^{j'}, C_{G^{j'}})}
\]

where \( d \) denotes the distance. Define the centroid distance as:

\[
C_d = d(C_{M_j}, C_{G^{j'}})
\]

and the deviation ratio as:

\[
R_d = \frac{D_{G^{j'}}}{D_{M_j}}
\]

then, the quality index \( Q^j \) is defined as:

\[
Q^j = \frac{2R_d^2}{(R_d^2)^2 + 1} [1 - \frac{C_d}{C_{\max}}]
\]

where \( C_{\max} = 2\sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2} \).  

The quality index \( Q \) has a dynamic range of \([0 1]\) with \( Q=1 \) associates with the highest quality. The highest value is reached when maximum similarity of the measured data and the ground truth data is achieved. This happens when the centroid distance is zero and the deviation ratio is 1. We also assume maximum similarity if both subsets of the pair \( (M, G') \) are empty, however if only one subset is empty zero value of \( Q \) is assumed.

Examples for values of \( N_m \) are 1, 8, 27, ... In general \( N_m = i^3 \) where \( i \) is a positive integer. The smaller the value of \( N_m \) the more tendency of \( Q \) to average errors. On the other hand, the greater the value of \( N_m \), the more sensitive the value of \( Q \) to outliers.

### 3 Experimental Results

A number of 36 images are acquired for a house object. Out of these images we select 4 images of orthogonal views at angles of 0, 90, 180, and 270. Two of these images are shown in Figure 1a and their silhouettes are shown in Figure 1b. A reference 3-D model for the house object is generated by the 3-D laser scanner. Only 10% of 3-D points of the scanner model is used to generate the silhouettes at the
same views of Figure 1b as shown in Figure 1c.

The search parameters are initialized with random values and a GA is applied on the given sets of silhouettes. The silhouettes of the 3-D scanner model resulted after running 100 generations/iterations of GA are shown in Figure 1d. The search parameters gained from GA step were applied to the next optimization step using the local search algorithm. 100% of scanner data are used in this step to help the algorithm getting accurate results. The results after running the NM method for 150 iterations are shown in Figure 1e.

The convergence of the search parameters is plotted for the six parameters in Figure 2. It is clear from the figure that GA provided good approximation for the parameters after less than 50 iteration and the NM method has refined the solution after less than 100 iterations.

To validate these results we plotted the ground truth data generated by the 3-D laser scanner before applying the RTS algorithm in the same plot with the 3-D reconstruction of the house object by the technique X applied to the 36 input images. The reconstruction by technique X plays no part in the registration process, but we plotted it to show the relative positions of the reference frames of the two reconstructions before applying the RTS technique as shown in Figure 3a. To show the alignment after applying RTS, the upper half of the technique X reconstruction and the lower half of the scanner reconstruction are plotted in Figure 3b. As a standard way of showing the registration results a mixed reconstruction is shown in Figure 3c where the brighter patches are generated by the scanner and the darker patches are generated by the technique X.

As a final comment on the algorithm performance, the accuracy of the registration depends on the number of distinct silhouettes used in the optimization phase. Of course, the greater the number the more accurate the results, however in the GA optimization step, the greater the number, the greater the time required to estimate the objective function specially when large population is assumed. That is why we used only 10% of the scanner data to reduce the run time in the GA optimization step.

We applied the proposed testing methodology to the reconstructions of two common 3-D reconstruction techniques; stereo and space carving. In both techniques we used the same set of input images of the house object.

Using \( N_m = 64 \), the quality index \( Q^j \) for each pair \( (M^j, G^j) \) is computed for the space carving data set. Figure 4a shows a bar graph for the quality index \( Q \) for each pair of data subsets and the histogram of the Q values is shown in Figure 4b. The Figure shows that almost 80% of the space carving reconstruction has Q values range from 0.8 to 1. This means that 80% space carving data set is in good match with the corresponding ground truth data and and almost 20% is in bad match. The percentage of bad match is increased to almost 25% in Figure 4c when the value of \( N_m \) is increased to 512 because the matched subsets became of smaller sizes then the errors are less averaged.

Analyzing the bad matched subsets, we found that they are mostly concentrated in the homogeneous areas of the house object. This is because the fattening problem of the basic space carving algorithm when the reconstructed object has large homogenous areas with smaller number of input images. This effect is shown in Figure 5a where the 3-D reconstruction is projected at a certain view and subtract the projected image from the input image at the same view. The reconstruction projected in Figure 5a is resulted from 9 input images. This effect is reduced when we increased the number of images to 12 as shown in Figure 5b.

The above experiment is repeated for the stereo reconstruction using \( N_m = 64 \). The bar graph and the histogram of Q are shown in Figure 6a, b. The stereo has quality index \( Q \) in the range 0.8-1 of 70% of its data subsets. As with space carving the homogenous areas in the 3-D object is responsible for low values of Q. However, in the stereo case these areas are not reconstructed hence many empty subsets are generated leaving quality index of almost zero. In addition, some errors in the stereo reconstruction are due to the mismatches.

In general, the space carving reconstruction has better results than stereo reconstruction even they both suffer from the same problem provided by the large homogenous areas as shown in Figure 6c however, with opposite responses.

Since the quality indices are computed for small patches of known positions in the 3-D space, then integration of different reconstruction techniques is possible by interchanging patches. The resulted reconstruction could be visualized as in Figure 3c.

4 Conclusions and Future Work

In this paper we propose a unified framework for the performance evaluation of 3-D reconstruction techniques. First, we propose a pre-evaluation technique that solves the problem of 3-D registration of measured and ground truth data by aligning their silhouettes. The algorithm does not assume any degree of dependance on the quality of the 3-D measured data, however it depends only on image silhouettes. The technique is simple, automatic, and efficient since it is not affected by the noisy reconstructions and does not assume any human intervention. In addition, we propose an evaluation methodology that aims at quantifying the performance of 3-D reconstruction techniques by extracting geometric features of uniformly distributed patches of both measured and ground truth data. The approach assigns a quality index for each patch based on the level of similarity of the geometric features. The approach has been applied to two 3-D reconstruction techniques; stereo and space
carving and it successfully localized common errors in both techniques.

In the future, we plan to perform extensive research on evaluation of different stereo and voxel based approaches based on the proposed framework. In addition, we plan to create a database that contains: input images, calibration and registration information, ground truth data for different objects and scenes and publish this database on our website for the public use.

References


Figure 1: (a) two out of four input images used by the RTS technique, (b) silhouettes of the input images in (a), (c) initial silhouettes from 3-D model generated by the scanner at same views as in (b) using 10% of scanner data, (d) final silhouettes after genetic algorithm application, (e) final silhouettes after application of the local search method starting with parameters generated by GA in (d) using 100% of scanner data.
Figure 2: Registration parameters: (a) rotation around X-axis, (b) rotation around Y-axis, (c) rotation around Z-axis, (d) translation in X-direction, (e) translation in Y-direction and (f) translation in Z-direction.

Figure 3: Registration: (a) unaligned 3-D reconstructions (b) after registration with the upper half reconstructed by technique X and the lower half is reconstructed by the 3-D by scanner, and (c) fusion of the reconstructions by the scanner and technique X to show the registration results.
Figure 4: (a) the bar graph of the quality index Q for different patches of space carving reconstruction at $N_m = 64$, (b) the histogram of the Q values in (a), and (c) the histogram of the Q values at $N_m = 512$.

Figure 5: A difference image between the reprojection of 3-D reconstruction by space carving and an input image at the same view, (a) given 9 input images, and (b) given 12 input images. Note the fattening problem in (a).

Figure 6: (a) the bar graph of the quality index Q for different patches of stereo reconstructions at $N_m = 64$, (b) the histogram of Q values in (a), and (c) the quality index values for both space carving and stereo at $N_m = 64$. 