AUTOMATIC 3D MAPPING USING MULTIPLE UNCALIBRATED CLOSE RANGE IMAGES

M. Rafiei, M. Saadatseresht

Department of Surveying Engineering, University of Tehran, Iran
(meysam.rafiei, msaadat)@ut.ac.ir

KEY WORDS: Close Range Photogrammetry, Automatic, 3D mapping, Euclidean Reconstruction, Uncalibrated Images

ABSTRACT:

Automatic three-dimensions modeling of the real world is an important research topic in the geomatics and computer vision fields for many years. By development of commercial digital cameras and modern image processing techniques, close range photogrammetry is vastly utilized in many fields such as structure measurements, topographic surveying, architectural and archeological surveying, etc. A non-contact photogrammetry provides methods to determine 3D locations of objects from two-dimensional (2D) images. Problem of estimating the locations of 3D points from multiple images, often involves simultaneously estimating both 3D geometry (structure) and camera pose (motion), it is commonly known as *structure from motion* (SfM). In this research a step by step approach to generate the 3D point cloud of a scene is considered. After taking images with a camera, we should detect corresponding points in each two views. Here an efficient SIFT method is used for image matching for large baselines. After that, we must retrieve the camera motion and 3D position of the matched feature points up to a projective transformation (projective reconstruction). Lacking additional information on the camera or the scene makes the parallel lines to be unparalleled. The results of SfM computation are much more useful if a metric reconstruction is obtained. Therefor multiple views Euclidean reconstruction applied and discussed. To refine and achieve the precise 3D points we use more general and useful approach, namely bundle adjustment. At the end two real cases have been considered to reconstruct (an excavation and a tower).

1. INTRODUCTION

There are many ways to mapping and obtain 3D models of the world around us and every approach has some advantage and disadvantage. In this regard some approaches are automatic, semi-automatic and some are non-automatic. Mostly non-automatic and semi-automatic manners are time-consuming, costly and need more than one operator. On the other hand, automatic approaches are very fast, simple and need one operator to perform. One of these methods is Close Range Photogrammetry. Digital close range photogrammetry is a low-cost technique for accurately measuring objects directly from digital images captured with a camera at close range.

Image-based modeling uses digital cameras and requires a mathematical formulation to transform 2D image coordinates into 3D information. Images contain all the useful information to form geometry and texture for a 3D modeling application (Luigi Barazzetti et al., 2010). Automatic techniques can be used to track the image features and solve it mathematically by using *Structure from Motion* (SfM) techniques, which refer to the computation of the camera stations and viewing directions (imaging configuration) and the 3D object points from at least two images (Alsadik et al., 2012).

Most of earlier studies in this field assume that the intrinsic parameters of the camera (focal length, image center and aspect ratio) are known. Computing camera motion in this case is a well-known problem in photogrammetry, called *relative orientation* (Atkinson, 1996). Some of them (Boufama et al., 1993) put some constraints on the reconstruction data in order to get reconstruction in the Euclidean space. Such constraints arise from knowledge of the scene: location of points, geometrical constraints on lines, etc. The paper of Barazzetti et al (2010) focused on calibrated camera. Cronk et al (2006) used coded targets for the calibration and orientation phase. Targets are automatically recognized, measured and labeled to solve the identification of the image correspondences. This solution becomes very useful and practical, but in many surveys targets cannot be used or stick at the object.

The aim of this paper is to report an automatic approach to generate the 3D point cloud of a scene with an uncalibrated camera. After taking images with one camera and constant focal length, we should detect corresponding points in each two views. A SIFT method is used for detecting and matching candidate features in pairs of images. Once the correspondence between features in different images has been established, we can directly recover the 3D structure of the scene up to a projective transformation. It will starts with the case of two views, and use the result to initializes a multiple-view algorithm. To reach the Euclidean model of the object, a linear transformation $H \in R^{4\times4}$ should be done. At the end, in order to refine the estimate obtained, a Euclidean bundle adjustment by minimizing the reprojection error is used.

2. THEORY

2.1 Image Matching

Features such as points and edges may change dramatically after image transformations, for example, after scaling and rotation. As a result, recent works have concentrated on detecting and describing image features that are invariant to these transformations. Scale Invariant Feature Transform (SIFT) is a feature-based image matching approach, which lessens the damaging effects of image transformations to a certain extent. Features extracted by SIFT are invariant to image scaling and rotation, and partially invariant to photometric changes. SIFT mainly covers 4 stages throughout the computation procedure as follows (D. Lowe, 2004):

1. Local extremum detection: first, use difference-of-Gaussian (DOG) to approximate Laplacian-of-Gaussian and build the image pyramid in scale space. Determine the keypoint candidates by local extremum detection.

2. Strip unstable keypoints: use the Taylor expansion of the scale-space function to reject those points that are not distinctive enough or are unsatisfactorily located near the edge.

3. Feature description: Local image gradients and orientations are computed around keypoints. A set of orientation, scale and location for each keypoint is used to represent it, which is significantly invariant to image transformations and luminance changes.

4. Feature matching: compute the feature descriptors in the target image in advance and store all the features in a shape-indexing feature database. To initiate the matching process for the new image, repeat steps 1-3 above and search for the most similar features in the database.

SIFT has been shown to be a valuable tool in 3D reconstruction. In this context (3D mapping) we face with large baseline in image acquisition stations. Some stations are near the object, the others are far from it probably. There is no smoothness in camera stations. In addition, rotation of camera is not like the aerial photography. Most the time the rotation matrix has large components. On the other hand, keypoints extracted by SIFT are highly distinctive so that they are invariant to image transformation and partially invariant to illumination and camera viewpoint changes. So we prefer to use SIFT method for feature extraction and feature matching steps. Figure 1 shows efficiency of the SIFT in image matching process.

2.2 The Camera Model

The projection of a point in space with coordinates $\boldsymbol{X}$ onto the image plane has (homogeneous) coordinates x' that satisfy the equation (1).

$$\lambda x' = K\Pi_0 gX = K[R,T]X, \tag{1}$$

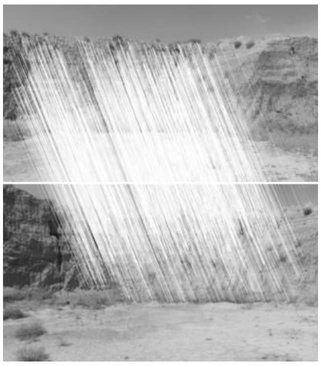


Figure 1. Efficiency of the SIFT in image matching

where $\Pi_0 = [I,0] \in \mathbb{R}^{3\times4}$ and $g \in SE(3)$ is the pose of the camera in the (chosen) world reference frame. In the equation above, the matrix K, which is define as

$$K = \begin{bmatrix} fs_x & s_\theta & o_x \\ 0 & fs_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{3\times3}, \tag{2}$$

describes "intrinsic" properties of the camera, such as the position of the optical center (o_x, o_y), the size of the pixel (s_x, s_y), its skew factor s_θ, and the focal length f. The matrix K is called the *intrinsic parameter matrix*, or simply *calibration matrix*, and it maps metric coordinates (units of meters) into image coordinates (units of pixels). In what follows, we denote pixel coordinates with a prime superscript x', whereas metric coordinates are indicated simply by $x(= K^{-1}x')$. The rigid-body motion $g = (R,T)$ represents the "extrinsic" properties of the camera, namely, its position and orientation relative to a chosen world reference frame. The parameters g are therefore called *extrinsic calibration parameters.*

2.3 The Fundamental Matrix *F*

The fundamental matrix is the algebraic representation of epipolar geometry. The fundamental matrix maps, or "transfers", a point x_1' in the first view to a vector $l_2\, F\, x_1' \in \mathbb{R}^3$ in the second view via

$$x_2'^T F\, x_1' = x_2'^T\; l_2 = 0.$$

In fact, the vector l_2 defines implicitly a line in the image plane as the collection of image points $\{x_2'\}$ that satisfy the equation

$$l_2^T\, x_2' = 0.$$

Given at least 8 correspondence, F can be estimated using the linear eight-point algorithm. If the number of correspondence was more than 8, a nonlinear algorithm should be used. Most of time, there are outliers in process of establishing correspondence and estimating F. These outliers eliminate in step of estimating the fundamental matrix simultaneously using RANSAC.

2.4 Projective Reconstruction

Once the correspondence between features in different images has been established, the 3D structure of the scene up to a projective transformation could be directly recovered. In the absence of any additional information, this is the best one can do. We start with the case of two views, and use the results to initialize a multiple-view algorithm (Yi Ma, et al., 2003).

2.4.1 Two View Initialization: The generic point $p \in \mathbb{E}^3$ has coordinates $X = [X, Y, Z]^{\mathrm{T}}$ relative to a fixed ("world") coordinate frame. Given two views of the scene related by a rigid-body motion *g = (R, T),* the 3D coordinate *X* and image measurements x_1' and x_2' are related by the camera projection matrices $\Pi_1, \Pi_2 \in \mathbb{R}^{3\times 4}$ in the following way:

$$\lambda_1 x_1' = \Pi_1 X, \quad \lambda_2 x_2' = \Pi_2 X \,,$$

$$\Pi_1 = [K, 0], \quad \Pi_2 = [KR, KT],$$

where $x' = [x', y', 1]^{\mathrm{T}}$ is measured (in pixel) and λ is an unknown scalar (the "projective depth" of point). The calibration matrix is unknown and has the general form of equation (2).

Recovering Projection Matrices and Projective Structure: Given the fundamental matrix *F* estimated via 8-point algorithm, there are several ways to decompose it in order obtain projection matrices and 3D structure from the two views. Since $F = \widehat{T'}KRK^{-1}$, all projection matrices $\Pi_p = [KRK^{-1} + T'v^T, v_4 T']$ yield the same fundamental matrix for any value of $v = [v_1, v_2, v_3]^T$ and v_4, and hence there is a four-parameter family of possible choices. One common choice, known as the *canonical decomposition*, has the following form

$$\Pi_{1p} = [I, 0], \qquad \Pi_2 = \left[\left(\widehat{T'}\right)^T F, T'\right], \;\; \lambda_1 x_1' = X_p \,,$$

$$\lambda_2 x_2' = \left(\widehat{T'}\right)^T F X_p + T'. \tag{3}$$

Now, different choices of v and v_4 result in different projection matrices Π_p, which in turn result in different projective coordinates X_p, and hence different reconstructions. Some of these reconstructions may be more "distorted" than others, in the sense of being farther from the "true" Euclidean reconstruction. In order to minimize the amount of projective distortion and obtain the initial reconstruction as close as possible to the Euclidean one, we can play with the choice of v and v_4, as suggested in (Beardsley et al., 1997). In practice, it is common to assume that the optical center is at the center of the image, that the focal length is roughly known (for instance from previous calibrations of the camera), and that the pixels are square with no skew. Therefore, one can start with a rough approximation of the intrinsic parameter matrix *K*, call it $\tilde{K}$. After doing so, we can choose $v \in \mathbb{R}^3$ and v_4 by requiring that the first block of the projection matrix be as close as possible to the rotation matrix between two views, $R \approx v_4\left(\widehat{T'}\right)^T F + T'v^T$. In case the actual rotation *R* between the views is small, we can start by choosing $\tilde{R} \approx I$, and solve linearly for v and v_4. In case of general rotation, one can still solve the equation $R \approx v_4\left(\widehat{T'}\right)^T F + T'v^T$ for v, provided a guess for the rotation $\tilde{R}$ is available. When the projection matrices, v and v_4 have chosen, the 3D structure can be recovered. Ideally, if guess of $\tilde{K}$ was accurate, all points should be visible; i.e. all estimated scales should be positive. If this is not the case, different values for the focal length can be tested until the majority of points have positive depth.

2.4.2 Multiple-View Reconstruction: When more than two views are available, they can be added one at a time or simultaneously. For the multiple-view setting we have

$$\lambda_i^j x_i^j = \Pi_i X^j, \quad i = 1,2,\ldots,m, \; j = 1,2,\ldots,n. \tag{4}$$

The matrix $\Pi_i = K_i \Pi_0 g_i$ is a 3 × 4 camera projection matrix that relates the *i*th (measured) image of the point *p* to its (unknown) 3D coordinates X^j with respect to the world reference frame. The goal of this step is to recover all the camera poses for the *m* views and the 3D structure of any point that appears in at least two views. For convenience, we will use the same notation $\Pi_i = [R_i, T_i]$ with $R_i \in \mathbb{R}^{3\times 3}$ and $T_i \in \mathbb{R}^3$.

The core of the multiple-view algorithm consists of exploiting the following equation, derived from the multiple-view rank conditions:

$$P_i \begin{bmatrix} R_i^s \\ T_i \end{bmatrix} = \begin{bmatrix} {x_1^1}^T \otimes \widehat{x_i^1} & \alpha^1 \widehat{x_i^1} \\ {x_1^2}^T \otimes \widehat{x_i^2} & \alpha^2 \widehat{x_i^2} \\ \vdots & \vdots \\ {x_1^n}^T \otimes \widehat{x_i^n} & \alpha^n \widehat{x_i^n} \end{bmatrix} \begin{bmatrix} R_i^s \\ T_i \end{bmatrix} = 0 \;\; \in \mathbb{R}^{3n}, \tag{5}$$

where ⊗ is the *Kronecker product.* Since $\alpha^j = 1/\lambda_1^j$, the inverse depth of X_1^j with respect to the first view, is known from the initialization stage from two views, the matrix $P_i \in \mathbb{R}^{3n\times 12}$ is of rank 11 if more than $n \geq 6$ points in general position are given, and the unknown motion parameters lie in the null space of P_i.

This leads to an algorithm which alternates between estimation of camera motion and 3D structure, exploiting multiple-view constraints available in all views. After the algorithm has converged, the camera motion is given by $[R_i, T_i]$, *i=2,3, . . . ,m,* and the depth of the points (with respect to the first camera frame) is given by $\lambda_1^j = 1/\alpha^j$, *j=1,2, . . . ,n.* The resulting projection matrices and the 3D structure obtained by the above iterative procedure can be then refined using a nonlinear optimization called *reprojection error*.

$$e_r = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|x_i^j - \pi(\Pi_i X^j)\right\|^2. \tag{6}$$

If the reprojection error is still large, the estimates can be refined using a nonlinear optimization procedure that simultaneously updates the estimates of both motion and structure parameters.

2.5 Upgrade From Projective To Euclidean Reconstruction

The projective reconstruction X_p obtained in the absence of calibration information, is related to the Euclidean structure X_e by a linear transformation $H \in \mathbb{R}^{4\times 4}$,

$$\Pi_{ip} \sim \Pi_{ie} H^{-1}, \;\; X_p \sim H X_e, \;\; i = 1,2,\ldots,m, \tag{7}$$

Where ~indicates equality up to a scale factor, $\Pi_{1p} = [I, 0]$ and *H* has the form

$$H = \begin{bmatrix} K_1 & 0 \\ -v^T K_1 & 1 \end{bmatrix} \;\; \in \mathbb{R}^{4\times 4}. \tag{8}$$

2.5.1 Stratification With The Absolute Quadric Constraint: According to the equation

$$\Pi_{ip}H \sim \Pi_{ie} = [K_iR_i, K_iT_i],$$

where we use $[R_i, T_i]$ to denote the Euclidean motion between the ith and the first camera frame. Since the last column gives three equations, but adds three unknowns, it is useless as far as providing constraints on H. Therefore, we can restrict our attention to the leftmost 3×3 block

$$\Pi_{ip}\begin{bmatrix} K_1 \\ -v^TK_1 \end{bmatrix} \sim K_iR_i. \qquad (9)$$

One can then eliminate the unknown rotation matrix R_i by multiplying both sides by their transpose:

$$\Pi_{ip}\begin{bmatrix} K_1K_1^T & -K_1K_1^Tv \\ -v^TK_1K_1^T & v^TK_1K_1^Tv \end{bmatrix}\Pi_{ip}^T \sim K_iK_i^T. \qquad (10)$$

If we define $S_i^{-1} \doteq K_iK_i^T \in \mathbb{R}^{3\times3}$, and

$$Q \doteq \begin{bmatrix} K_1K_1^T & -K_1K_1^Tv \\ -v^TK_1K_1^T & v^TK_1K_1^Tv \end{bmatrix} \in \mathbb{R}^{4\times4}, \qquad (11)$$

then we obtain the *absolute quadric constraint* as bellow

$$\Pi_{ip}Q\Pi_{ip}^T \sim S_i^{-1}. \qquad (12)$$

If we assume that K is constant, so that $K_i = K$ for all i, then we can minimize the angle between the vectors composing the matrices on the left-hand side and those on the right-hand side with respect to the unknowns, K and v, using for instance a gradient descent procedure. Alternatively, it could first estimate Q and K_i from this equation by ignoring its internal structure; then, H and K can be extracted from Q, and subsequently the recovered structure and motion can be upgraded to Euclidean. Figure 2 shows the Euclidean reconstruction of a projective form.

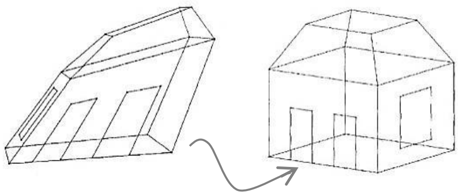


Figure 2. Upgrade from projective to Euclidean reconstruction

2.6 Euclidean Bundle Adjustment

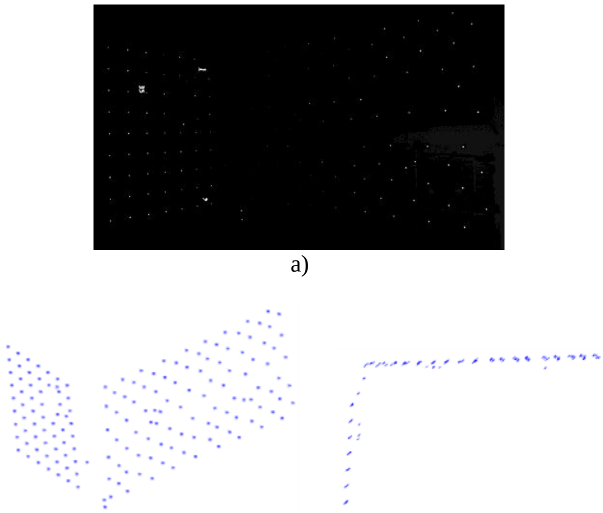
In order to refine the estimate obtained so far, we can set up an iterative optimization, known as *Euclidean bundle adjustment*, by minimizing the reprojection error as in Section 2.4.2, but this time with respect to all and only the unknown Euclidean parameters: structure, motion, and calibration. The reprojection error is still given by

$$e_r = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|x_i^j - \pi(K_i(R_iX^j + T_i))\right\|^2. \qquad (13)$$

However, this time the parameters are given by $\xi \doteq \{K_i, \omega_i, T_i, X^j\}$, where ω_i are the exponential coordinates of rotation $R_i = e^{\widehat{\omega}_i}$ and are computed via Rodrigues' formula. The total dimension of the parameter space is $5 + 6(m - 1) + 3n$ for m views and n points (Yi Ma, et al., 2003).

3. IMPLEMENTATION

At first, we created a test field to evaluate of our approach. The test field was created from some retro-reflective targets to compare our technique with the Australis software. We used this kind of test field because it was a simple and good way to test the 3D metric reconstruction. Figure 3 shows the test field and its 3D model.



a)

b)

Figure 3. a) The test field b) Euclidean reconstruction of the test field

To implement our approach in real word, we consider two cases to reconstruct: an excavation and a historical tower. In this study, the Canon SX230 digital camera (a non-metric digital camera with Manual setting to fix the focal length) has been used (Figure 4).

Figure 4. Canon SX230 digital camera

3.1 Case Studies

Excavation: One of the most important applications of 3D mapping with uncalibrated images is volume computation. This is a very fast manner to compute the volume of excavation or embankment of soil in many projects. Since, we considered an excavation 3D modeling as shown in Figure 5 to evaluate our approach.

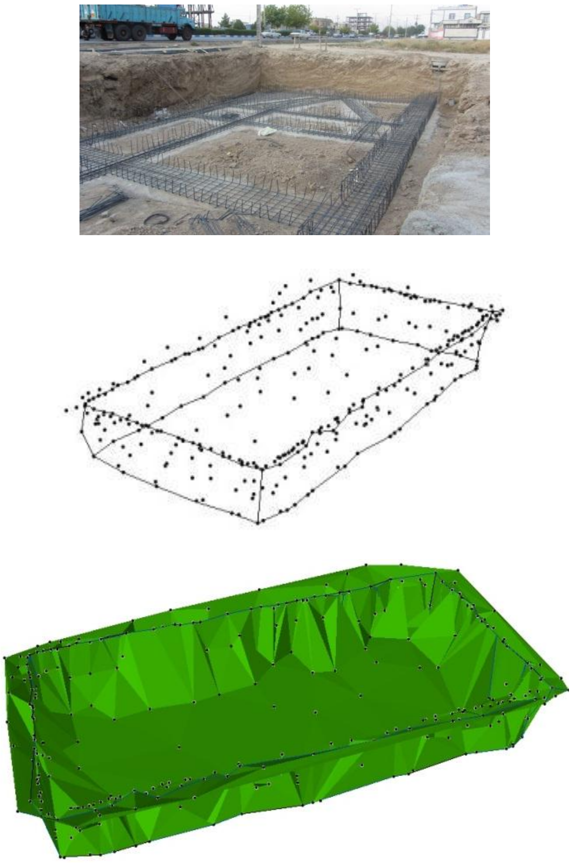
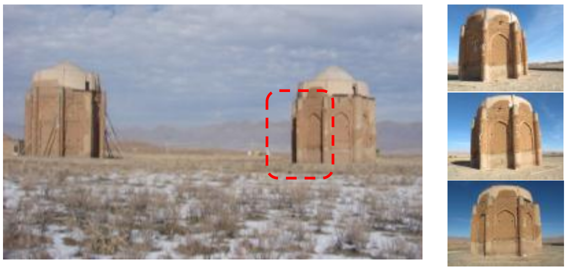


Figure 5. The excavation, its 3D point clouds and model

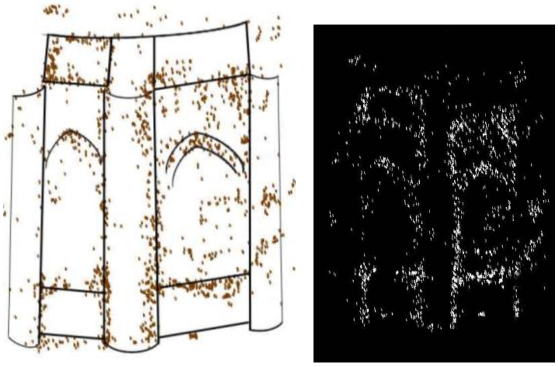
Area of the excavation field was about 400m^2. The 22 consequence images have been taken to cover all the area.

Kharraqan Towers: The need for documentation of historic buildings is greatly admitted by many experts in numerous literatures. One step in documentation is surveying or 3D modeling. In this regard, digital close range photogrammetry shows a high proficiency and it is a low-cost technique to collects 3D information about an object. The two Kharraqan Tomb Towers are located in 33km west of Ab-e Garm town, in Qazvin Province in Iran. These two towers are particularly notable for their vivid external decoration, which classes them amongst the finest decorated brick monuments found in Iran (Briseghella & Kalhor, 2007).

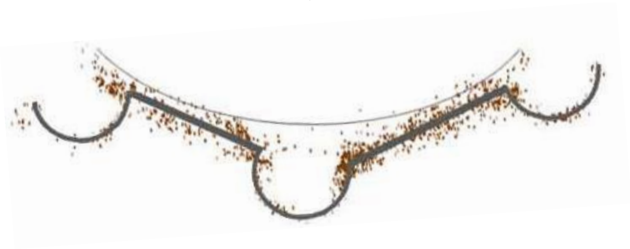
We consider just two sides of west tower for modeling. The model of two faces of this octagon tower has been formed with 6 images. A superficial network is designed to reach the minimum number of at least three cameras for each point (Fraser, 1989). Figure 6 shows the efficiency of this technique to reconstruct the historical tower.



a)



b)



c)

Figure 6. a)Photos of Tower(s) b) Its 3D point clouds with an sketchy model c) Plan of two faces

4. RESULT AND DISCUSSION

With the use of the test field, a comparison between our technique and Australis 3D modeling has been done. Table 1 shows the differences between each direction.

Direction	X	Y	Z
RMS (mm)	**1.3**	**1.5**	**3**

Table 1. Accuracy assessment between our technique and Autralis (Z axis is in direction of first camera's focal length)

When experimental result examined, photogrammetric technique is efficient for fast 3D mapping. The first case study modeling has been showed about 8cm differences in a one large length at a huge excavation. It is so better in Kharraqan Tower reconstruction; the difference was about 3cm.

This method has been shown that by generating dense point clouds, it is more effective than classical surveying in such volume calculation. In addition, image based modeling could approach 97.3 % ratio to the real value. On the other hand, it is a very low cost and fast technique. It reduces the land work time and usually needs one people to work, but classical land surveying needs at least 3 people to work. Consequently, photogrammetric methods have 33.33% cost advantage in field works when compared with the classical method (M.Yakar et al., 2008).

There is an experimental relationship between distance to object and camera stations distance. If distance to object is large, it is better to have a large camera station baseline. It forms a strong network. It is very important to peruse the object before imaging. If the object or area is large, we have to take more images and it increases the computational process (in the case of partial imaging). Moreover, it leads to block error in bundle block adjustment if the images don't have an adequate overlap (more than 60%). This is very important to select an appropriate digital camera proportional to the project. If the object is big or distance to object is large, it should be to use a camera with high resolution. It is suggested that for 3D reconstruction projects it is better to use SLR digital cameras.

5. CONCLUSION

By the development of off-the-shelf digital cameras and modern image processing techniques, digital close range photogrammetry is vastly utilized in many fields such as engineering structure measurements, topographic and archeological surveying, etc. Both cheap and fast features lead to vast use of this method.

In this research, we consider an approach to reach a three-dimensional Euclidean model of an object or a scene by means of an automatic manner. By the help of perspective projection, it is possible to goes from 2D images to 3D features. Problem of estimating the locations of 3D points from multiple images commonly known as *structure from motion* (SfM). We used an efficient SIFT method to solve the correspondence problem, which it is a basic step in SfM. After that to reach the metric reconstruction, we passed the projective transformation. At the end a nonlinear bundle adjustment was applied to refine the estimated obtained.

To implement the mentioned technique, we started with a test field to evaluate our technique. This kind of test field has been used to compare our approach with Australis software. The result show acceptable differences and confirm our approach for 3D metric modeling. Then for testing above, we consider an excavation and a historical tower to reconstruct. The excavation case was elected because of widely usage of this approach in such fields, and to show the performance of this method to documentation of historical places, the Kharraqan Tomb Towers was selected. To cover the two cases, respectively, 22 and 6 images have been taken for excavation and historical tower.

REFERENCES

Atkinson, K.B., 1996. *Close Range Photogrammetry and Machine Vision*, Whittles Publishing, Scotland.

Barazzetti, L., Scaioni, M., Remondino, F., 2010. Orientation and 3D modeling from markerless terrestrial images: combining accuracy with automation. *The Photogrammetric Record* 25, pp.356-381.

B. Boufama R. Mohr F. Veillon, 1993. Euclidean Constraints for Uncalibrated Reconstruction. IEEE.

Briseghella L. and M. Kalhor, 2007. The Seismic Rehabilitation of Kharraqan Tomb Towers. *The 5th international conference on seismology and earthquake engineering*, Tehran, Iran.

B. S. Alsadik, M. Gerke, G. Vosselman, 2012. OPTIMAL CAMERA NETWORK DESIGN FOR 3D MODELING OF CULTURAL HERITAGE. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume I-3,* Melbourne, Australia.

Cronk, S., Fraser, C. and Hanley, H., 2006. Automated metric calibration of colour digital cameras. *Photogrammetric Record,* 21(116): 355–372.

Fraser, C.S., 1989. *Non Topographic Photogrammetry*, 2nd edition ed. Edwards Brothers Inc.

Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision.*

M.Yakar, H.M. Yilmaz, 2008. USING IN VOLUME COMPUTING OF DIGITAL CLOSE RANGE PHOTOGRAMMETRY, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXVII. Part B3b. Beijing.

YiMa, Stefano Soatto, J ana Kosecka, S. Shankar Sastry, 2004. *An Invitation to 3D vision*. Springer, pp. 375-403.