VNU Journal of Science, Mathematics - Physics 23 (2007) 210-220

Toward building 3D model of Vietnam National University, Hanoi (VNU) from video sequences

Trung Kien Dang, The Duy Bui*

College of Technology, VNU
144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

Received 9 Jun 2006; received in revised form 30 Jun 2006

Abstract. 3D models are getting more and more attention from the research community. The application potential of 3D models is enormous, especially in creating virtual environments. In Vietnam National University - Hanoi, there is a need for a test-bed 3D environment for research in virtual reality and advance learning techniques. This need raises a very good motivation for the research of 3D reconstruction. In this paper, we present our work toward the creating of a 3D model of Vietnam National University - Hanoi automatically from image sequences. We use the reconstruction process proposed in [1], which consists of four main steps: Feature Detection and Matching, Structure and Motion Recovery, Stereo Mapping, and Modeling. Moreover, we develop a new technique for the structure update step. By applying proper transformation on the input of the step, we have produced a new simple but effective technique which has not been considered before in the literature.

1. Introduction

Recently, 3D models are getting more and more attention from the research community. The application potential of 3D models is enormous, especially in creating virtual environments. A 3D model of a museum allows the user to visit the museum "virtually" just by sitting in front of the computer and clicking mouse. A security officer of a university can check the classroom "virtually" through the computer. This is the result of mixing real information from security camera with a 3D model. In order to build 3D models, the tradition is normally used, in which technicians builds the 3D models manually and then apply the texture on these models. This method requires enormous manual effort. With five technicians, it may require three to six months to build a 3D model. When a change is needed, manual effort is required again. The model may even have to rebuild from the scratch. A new approach is investigated to reduce the human effort is to build 3D models automatically from video sequences.

In Vietnam National University, Hanoi, there is a need for a test-bed 3D environment for research in virtual reality and advance learning techniques. This need raises a very good motivation for the research of 3D reconstruction. Again, the question is how to create a 3D model of Vietnam National University - Hanoi with the least human effort.

* Corresponding author. E-mail: duybt@vnu.edu.vn

In this paper, we present our work toward the creating of a 3D model of Vietnam National University, Hanoi automatically from image sequences. Among many proposed methods (e.g. [2, 3, 4, 5]) we chose the framework proposed in [1] because of its completeness and practicality. The reconstruction described in [1] consists of four main steps: Feature Detection and Matching, Structure and Motion Recovery, Stereo Mapping, and Modeling. Moreover, we develop a new technique for the structure update step. By applying proper transformation on the input of the step, we have produced a new simple but effective technique which has not been considered before in the literature.

Section 2 gives an overview of the 3D reconstruction process that we use to build the 3D model. We then propose our technique for the structure update step in Section 3. We then show the experiments that we have done to show the effectiveness of our technique in Section 4.

2. The 3D reconstruction process

We follow the 3D reconstruction process implemented in [1], which is illustrated in Figure 1. The process consists of four main steps: Feature Detection and Matching, Structure and Motion Recovery, Stereo Mapping, and Modeling. These steps will now be discussed in more details.

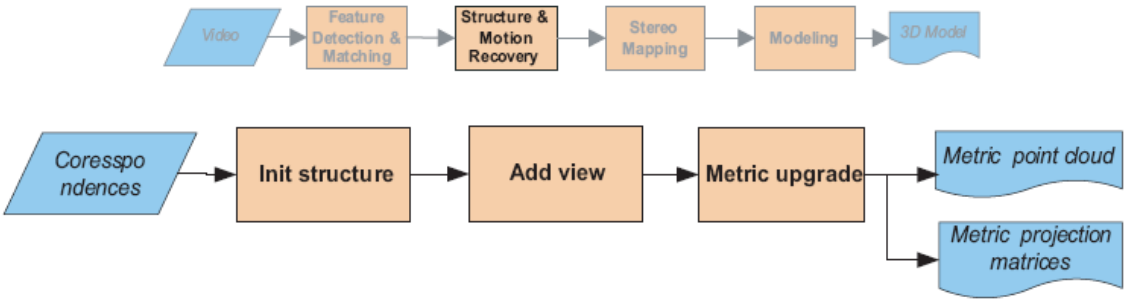

Fig. 1. Main tasks of 3D reconstruction with detail of the Structure and Motion recovery step.

2.1. Feature Detection and Matching

The first step involves in relating different images from a collection of images or a video sequence to each other. In order to determine the geometric relationship (or multi-view constraints) between images, it requires a number of corresponding feature points. Feature points are point that can be differentiated from its neighboring image points so that it can be matched uniquely with a corresponding point in another image. These features points are then used to compute the multi-view constraints, which corresponds to the epipolar geometry and is mathematically expressed by the fundamental matrix. This fundamental matrix can be found by solving 8 linear equations. Hartley has pointed out that normalizing the image coordinates before solving the linear equations would reduce the error caused by the difference by several orders of magnitude between columns in linear equations. The transformation is done by transforming the image center to the origin and scaling the images so that the coordinates have a standard deviation of unity.

2.2. Structure and Motion Recovery

At this step, the structure of the scene and the motion of the camera is retrieved using the relation between the views and the correspondences between the features. Among the 4 main steps of the 3D reconstruction it is extremely important for the accuracy of the final model since it defines the "skeleton" of the model. The process starts with creating an initial reconstruction frame with two images. Two images suitable for the initialization process are selected so that they are not too close to each other on the one hand and there are sufficient features matched between these two images on the other hand. The reconstruction frame is then refined and extended each time a new view (image) is added. The pose of the camera for each new view is estimated so that views that have no common features with the reference views also becomes possible. A projective bundle adjustment can be used to refine the structure and motion after it is determined for the whole sequence of images. This is recommended to be done with a global minimization step. Nevertheless, the reconstruction so far is only determined up to an arbitrary projective transformation. This is not sufficient enough for visualization. Therefore, the reconstruction need to be upgraded to a metric one, which is done by a process called self-calibration which imposes some constraints on the intrinsic camera parameters. Finally, in order to obtain an optimal estimation of the structure and motion, a metric bundle adjustment is used.

2.3. Stereo Mapping

At this stage, the methods developed for calibrated structure from motion algorithms can be used as the camera calibration has been done for all viewpoints of the sequence. Although the feature tracking algorithm has produced a sparse surface model, this is not sufficient to reconstruct geometrically correct and visually acceptable surface models. A dense disparity matching step is required to solve this problem. The dense disparity matching is done by exploiting additional geometrical constraints which is performed in several steps: (i) image pairs are rectified so that epipolar lines coinciding with the image scan lines which reduces the correspondence search to a matching of the image points along each image scan-line; (ii) disparity maps are computed through a stereo matching algorithm; (iii) a multi-view approach integrates the results obtained from several view pairs by fusing all independent estimates into a common 3D model.

2.4. 3D Modeling

To reduce geometric complexity, a 3D surface is approximated to the point cloud generated by previous steps. This step also tailors the model so it can be displayed by a visualization system.

3. Coordinate normalization for structure update

In this section we motivate and present our normalization technique for structure update and its relation to others.

3.1. Coordinate Normalization

The inputs for the metric upgrade are *canonical representations* [6] of at least four views' projection matrices. A practical approach was proposed in [1]. First the fundamental matrix of the two

initial views is decomposed into two projection matrices. The first 3D points, i.e. the initial projective structure, are than recovered by finding the intersections of back-projected rays, a triangulation process. Then projections of the initial 3D points on a new view are found to establish the equation system which allows adding that view to the projective structure. The view adding process is iterative and is called structure update.

For each 3D to 2D correspondence (X, x), from the projection equation x = PX, we have two equations to compute the projection matrix of the new view.

$$\begin{bmatrix} -X & 0^{(4)} & x_0 X \\ 0^{(4)} & -X & x_1 X \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = 0 \quad (1)$$

where p_i (i = 1,2,3) are row vectors of the new view's projection matrix P_{new}.

Since we have 12 unknowns (recall that P is a 3×4 matrix), at least six correspondences are required to solve the problem.

Based on our real data observation we assume the X_i (i = 0,1,2) are about 10 and similar to [7] x_i (i=0,1) are about 100. Let A denote the coefficient matrix. For the assumed values the corresponding entries of a row of A are of the following magnitude r(−10, −10, −10, −1, 0, 0, 0, 0, 10^3, 10^3, 10^3, 10^2). The entries of A^TA are approximated rr^T = (10^2, 10^2, 10^2, 1, 0, 0, 0, 0, 10^6, 10^6, 10^6, 10^4). That means the values of the entries range from 0 to 10^8. For an intuitive stability analysis, we can assume that the diagonal of A^TA is (10^6, ..., 1).

Let λ_i denote an eigenvalue of the matrix ($\lambda_i \le \lambda_j$, i < j), and $M_{12} = A^TA$. We wish to estimate the condition number $\kappa = \lambda_1(M_{12})/ \lambda_{12}(M_{12})$. Given that A^TA is symmetric, and using the *Interlacing Property* [8], we can deduce two facts: (i) the largest eigenvalue of M_{12} is no less than the largest diagonal entry $\lambda_1(M_{12}) \ge 10^8$, (ii) and the smallest one $\lambda_{12}(M_{12}) \le \lambda_1(M_1) = 1$. Thus the condition number of M_{12} is $\kappa = \lambda_1/ \lambda_{12} \ge 10^6$, which is a very large number. Here implies that noise can have significant impact.

Coordinate normalization before the structure update can reduce the condition number. Because we must maintain the consistency over the projection matrix chain, the transformation must be the same for every frame. Hence we have to find a transformation based on the expected values of the data rather than specific values. The assumption we used here is that the feature points are distributed uniformly around images' center and that the fixed frames' size is known.

So with the feature points are distributed around the image center, we first need a transformation to make the image center the origin:

$$T_T = \begin{bmatrix} 1 & 0 & -h/2 \\ & 1 & -h/2 \\ & & 1 \end{bmatrix} \quad (2)$$

in which w and h are the frames' width and height respectively.

After that, to equal the magnitudes of homogeneous coordinates, the scaling transformation should reduce the average distance of feature points to their centroid. For simplicity we use the following transformation to get that effect:

$$T_S = \begin{bmatrix} \frac{k}{\sqrt{w^2 + h^2}} & 0 & 0 \\ & \frac{k}{\sqrt{w^2 + h^2}} & 0 \\ & & 1 \end{bmatrix} \quad (3)$$

in which *k* is a scalar. In our experiments it is set to $\sqrt{2}$ as we want to limit coordinates to a (1, 1) rectangle. Consequently, 3D points of the projective structure are scaled to seemingly fit into a unit box.

Together the transformation is:

$$T_N = T_S T_T \quad (4)$$

This transformation will minimize the effect of unbalanced coordinate magnitudes. Below we will explain how to apply it in more detail.

3.2. How to apply the technique

In this sub-section we explain more of how to apply the techniques and its relation to other methods. Also we show how to adjust others once our technique is applied.

Although the technique is to improve the structure update, it must be applied before the structure initialization for two reasons: (i) to keep the added views' consistent to initial views, (ii) and to reduce the unbalance among elements of initial 3D points. As it is applied before the structure initialization, the threshold to decide on outliers in the robust fundamental matrix computation must be adjusted.

The normalization to prepare for the metric upgrade [1] should also be adjusted. There is no need to translate the origin to the center of the images anymore. The *w* and *h* are new scaled dimensions of the picture. Thus K_N should now be:

$$K'_N = \begin{bmatrix} w' + h' & 0 & 0 \\ & w' + h' & 0 \\ & & 1 \end{bmatrix} \quad (5)$$

Table 1 gives the outline of the order of the steps of the structure and motion recovery with the new normalization technique.

4. Experiments and discussion

In this section we give the results of our technique on synthetic and real data. The synthetic experiment setup is based on some related work. The real data include one traditional sequence in 3D reconstruction and two others from our experimental video for the application we are aiming at the reconstruction of Vietnam National University, Hanoi.

Table 1. Normalizations in structure and motion recovery.

1. **Projective structure initiation**
   (a) ***Image coordinate normalization with*** T_N***:*** $x' = T_N x$
   (b) Compute F with adjusted outlier threshold. Decompose projection matrices from F.
   (c) Triangulate to initiate the projective 3D point cloud.
   (d) Add views.
2. **Metric upgrade.**
   (a) Projection matrices normalization with adjusted K'_N: $P_N = K'^{-1}_N P$.
   (b) Compute metric upgrade homography. Upgrade the structure and motion.

4.1. Synthetic data

Synthetic input used is a random 3D point cloud uniformly distributed within a cubic. To their projections onto frames and the principal point with zero mean Gaussian error of standard deviation of 0.5 and 0.1 point is respectively added. The setup is based on the setup of experiments in [9, 1] and the assumption that the image point error is mainly caused by the digitization. The result is the average of 100 runs.

Evaluation criteria are twofold. The condition number graph shows how our technique reduces the sensitity of the solution to input noise. The reprojection error is used to evaluate the actual improvement. Since the frames are scaled down by normalization, the absolute geometric error no longer reflects the improvement. Thus to measure the geometric improvement, we convert the reprojection error back to the original coordinate scale using this equation.

$$\mathrm{err} = \frac{|\,\mathrm{PX} - \mathrm{x}\,|}{\text{scale factor}} \tag{6}$$

where the scale factor is 1.0 in the non-normalized case and $\sqrt{2}/\sqrt{\mathrm{w}^2 + \mathrm{h}^2}$ in the normalized case.

Figure 2 shows the average condition number on a logarithmic scale with respect to the number of points used to add a new view. Note that the condition number without normalization is about 10^7, close to our estimate in the previous section. It is reduced about 10^4 to 10^5 times. This helps to achieve a better result as showed in Figure 3. The reprojection error is reduced from about 1.0 to less than 0.01 pixel.

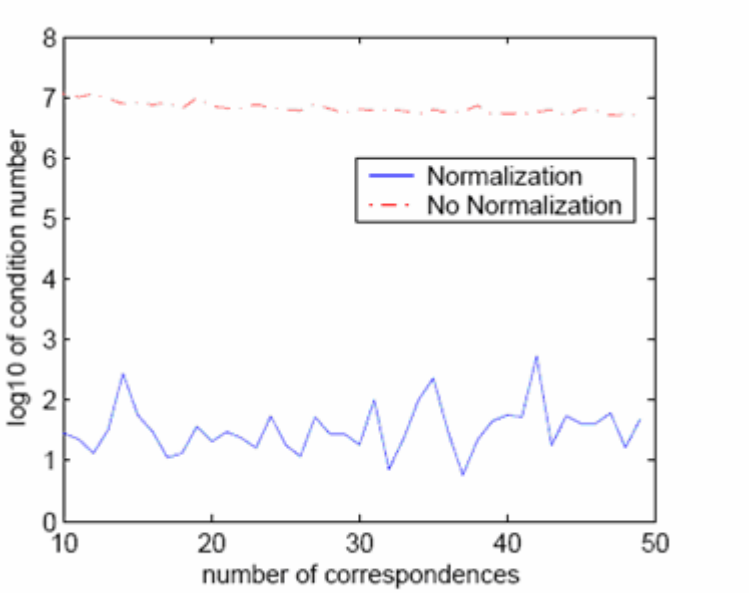

Fig. 2. Log10 of condition number vs. number of correspondences.

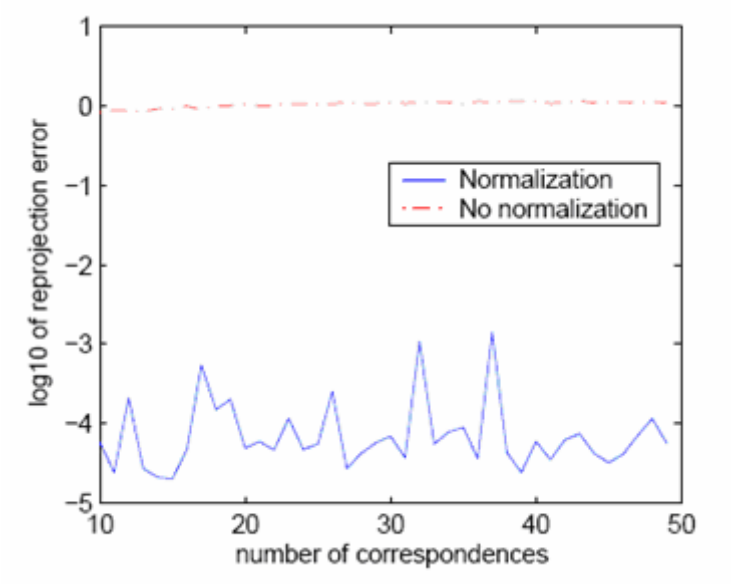

Fig. 3. Log10 of reprojection number vs. number of correspondences.

To see the relation between input noise and output error we fix the number of correspondences at 30 and vary the input noise standard deviation from 0.2 to 1.6 pixels. Figure 4 and 5 show the dependency of the condition number and the reprojection error on the input error. As the input noise

increases the reprojection error without normalization increases, with normalization the error stays much smaller.

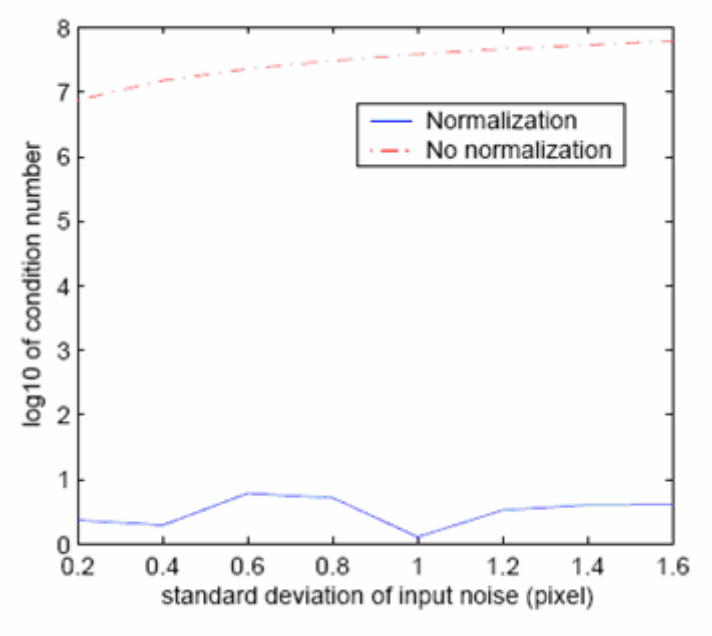

Fig. 4. Log10 of condition number vs. input noise.

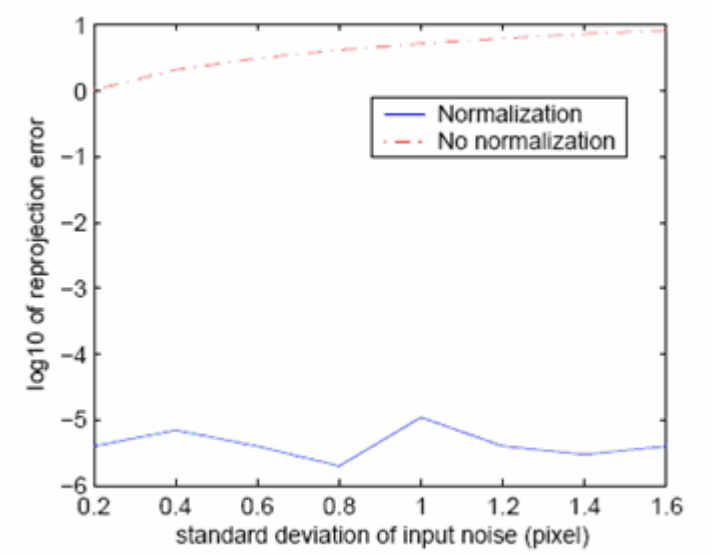

Fig. 5. Log10 of reprojection number vs. input noise.

The results are however not always stable in the normalized case. It is probably because in some cases the assumptions do not hold thus the condition number and consequently the error is not reduced as expected. We will have to examine those cases further.

4.2. Real data

The new technique is tested with real images of Vietnam National University, Hanoi (see Figure 6). In addition compared to the process explained in Table 1 RANSAC is used in the structure update in order to reject outliers that cannot be rejected when computing F. In most of the cases the result is similar to the synthetic experiment's result.

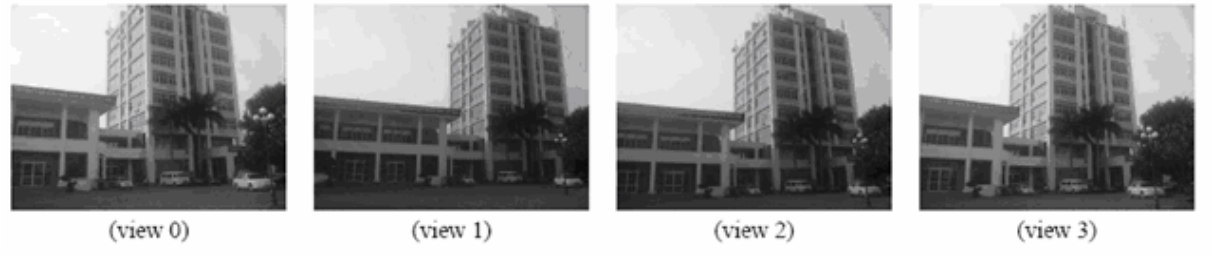

Fig. 6. Experimental image sequences of Vietnam National University, Hanoi.

In this sequence we used four frames, two to initiate the structure and two for added views, in order to have enough views for metric upgrade [9]. Figure 7 shows the feature points detected on the image sequences, while Figure 8 shows how these features points are matched.

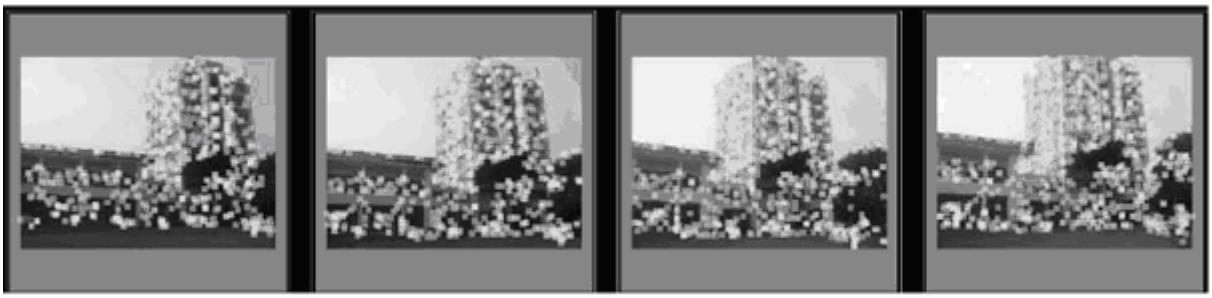

Fig. 7. Feature points detected on the image sequences of Vietnam National University, Hanoi by SFTF [10].

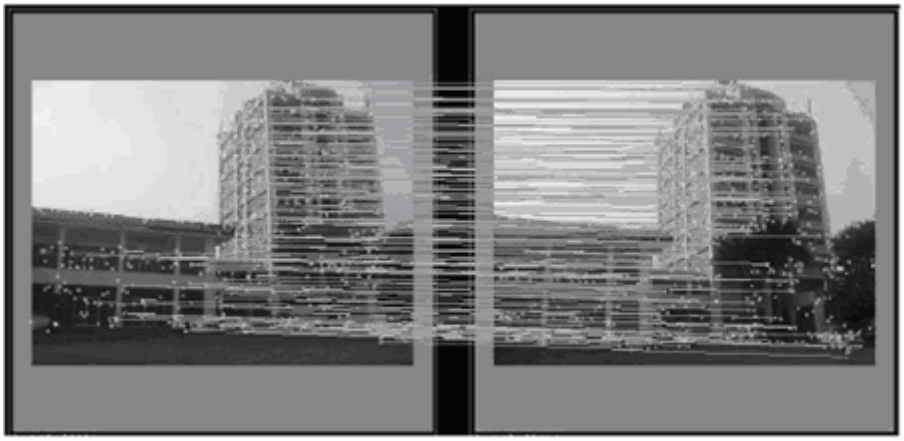

Fig. 8. Feature points on the image sequences are matched

The condition number and reprojection error are given in table 2 and 3 respectively. As can be seen from the table, the result shows that technique has improved the condition number and reprojection error for the image sequence. This is rather close to the synthetic result.

Table 2. Condition number with/without the normalization.

Seq.	Norm	Non Norm
View 2	3053.945774	12733610.731249
View 3	4745.445946	7462514.512543

Table 3. Reprojection error with/without the normalization

Seq.	Norm	Non Norm
View 2	0.000472	0.314433
View 3	0.000367	0.963279

After the 3D reconstruction process, the generated point cloud is shown in Figure 9. Using polar rectification [11] and a simple dynamic programming stereomapping we generate the final 3D model of Vietnam National University, Hanoi that is shown in Figure 10. Due to the simplicity of the stereo mapping algorithm, detail of the model is lost. In future, to improve the quality we will try to use higher quality images as well as apply more sophisticated algorithms (e.g. [12, 13]).

Fig. 9. Point cloud generate for the 3D model of Vietnam National University, Hanoi.

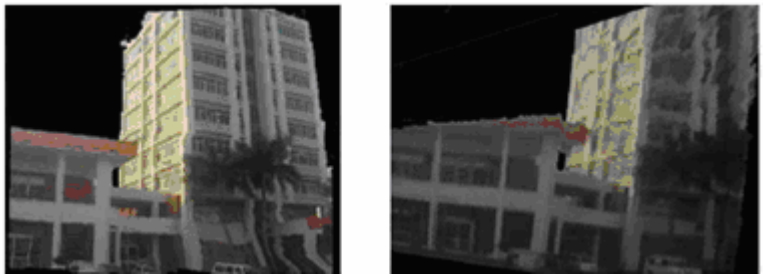

Fig. 10. 3D model of Vietnam National University, Hanoi.

5. Conclusion

We presented in this paper our work toward the creating of a 3D model of Vietnam National University, Hanoi automatically from image sequences. Using a reconstruction process proposed in [1], we have generate a 3D model of Vietnam National University, Hanoi with a fair overall quality. The quality of the 3D model is improved by a new technique that we developed for the structure update step. In the future we want to improve the reconstruction process more in order to have a more detailed and accurate 3D model.

References

[1] M. Pollefeys, Visual modeling with a hand-held camera, *International Journal of Computer Vision* 59 (2004) 207.
[2] R. Hartley, A. Zisserman, *Multiple view geometry in computer vision – 2nd edition.* Cambridge University Press, 2004.
[3] J. Ponce, K. McHenry, T. Papadopoulo, M. Teillaud, B. Triggs. On the absolute quadratic complex and its application to autocalibration. *IEEE Conference on Computer Vision and Pattern Recognition* (2005) 780.
[4] M. Han, T. Kanade, A perspective factorization method for euclidean reconstruction with uncalibrated cameras, *Journal of Visualization and Computer Animation* 13 (2002) 211.
[5] M. Ming-Yuen Chang, K. Hong Wong, Model reconstruction and pose acquisition using extended Lowe's method. *IEEE Transaction of Multimedia* 7 (2005) 253.
[6] Q.T. Luong, T. Vieville, Canonical representations for the geometries of multiple projective views, *Computer Vision and Image Understanding* 64 (1996) 193.
[7] R.I. Hartley, In defense of the eight-point algorithm, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (1997) 580.
[8] G.H. Golub, C.F. Van Loan, *Matrix Computation – 3nd edition,* Johns Hopkins University Press, 1996.
[9] M. Pollefeys, R.Koch, L.V. Gool, Self calibration and metric reconstruction in spite of varying and unknown intrinsic camera parameters, *IEEE International Conference on Computer Vision* (1998) 90.
[10] D.G. Lowe, Distinctive image features from scale invariant keypoints, *International Journal of Computer Vision* 60 (2004) 91.
[11] M. Pollefeys, R. Koch, L.V. Gool, A simple and efficient rectification method for general motion, *International Conference on Computer Vision* (1999) 496.
[12] J. Sun, Y. Li, S.B. Kang, H.Y. Shum, Symmetric stereo matching for occlusion handling, *International Conference on Computer Vision and Pattern Recognition* (2005) 399.
[13] Y. Wei, L. Quan, Asymmetrical occlusion handling using graph cut for multi-view stereo, *IEEE Conference on Computer Vision and Pattern Recognition* (2005) 902.