MAP Model for Large-scale 3D Reconstruction and Coarse Matching for Unordered Wide-baseline Photos

Xiuyuan Zeng, Qing Wang, Jiong Xu
School of Computer Science and Engineering
Northwestern Polytechnical University
Xi'an 710072, P. R. China
qwang@nwpu.edu.cn

Abstract

In this paper we presented a novel idea for large-scale 3D scene reconstruction and annealing based image grouping algorithm for unordered wide-baseline photos. Firstly, an alternative maximum a posterior (MAP) model which can easily incorporate image clustering prior knowledge is proposed. Second, an efficient annealing clustering algorithm is developed for organizing photos into clusters by calculating matching number of invariant features. Thirdly, we analyze the time complexity and efficiency of the proposed approach. Finally a series of experiments are performed on the real image data and synthetic data. The experimental result shows that the MAP model and relative annealing algorithm are efficient enough to tackle the large-scale 3D reconstruction problem, and it can be extended to solve other similar SFM parameters estimation problem as well.

1 Introduction

Large-scale 3D reconstruction, as a challenging issue in computer vision application, has drawn considerable attentions in last decade, and a lot of efforts have been devoted to develop efficient approaches for recovering high-quality 3D scene models from a large set of unordered and wide-baseline images, which are taken from widely separated viewpoints. The key problem of large-scale 3D reconstruction is the wide-baseline stereo (WBS) matching [1]. The procedure of WBS matching includes two steps: first, one is to find and extract local features from images by invariant descriptor, and then it is necessary to build up accurate correspondence between them. However, WBS matching is much more complicated and difficult than traditional small-baseline matching, for it has to tackle large deformation and affine wrapping due to the large changes over viewpoints.

So far, most of the literatures and existed methods [2-4] are focus on the two-view WBS matching. For multi-view WBS matching in large-scale 3D scene reconstruction, the most simple and direct method is to treat it as a series of two-view (tri-view) WBS matching, like the Photo Tourism system presented in [5]. The system computed the correspondences between each possible images pair and jointed the reconstruction

BMVC 2008 doi:10.5244/C.22.34

results together to finally recovery the completely 3D model. Unfortunately, the computational cost of Photo Tourism system cannot be acceptable if the number of image is large, especially when the images are unordered and wide-baseline.

Dellaert et al. [6-7] presented a novel Maximum likelihood (ML) statistical model for 3D reconstruction and employ MCEM algorithm to recover the 3D scene structure and camera motion parameters together. In each iteration of the MCEM algorithm, the correspondence is regarded as hidden variable and represented as a probability distribution type instead of single best correspondence vector, and it is estimated by an efficient Markov chain Monte Carlo (MCMC) sampling method [8] in each E-step. Moreover in recent work, Dellaert et al. [9] presented an out-of-core bundle adjustment algorithm for generating large-scale 3D reconstructions, which can be used as M-step of EM algorithm. However, a good global reconstruction result is hard to obtain since the ML statistical model is too simple to describe the unordered and wide-baseline characteristic of images in large-scale 3D reconstruction. As a result, the computational cost is still high and the MCMC sampling method in MCEM algorithm may fail to estimate a good result for 3D structure and camera motion parameters.

Alternatively, it is possible to carry out coarse matching before 3D scene reconstruction. The main idea of coarse matching is to find the multi-view feature correspondence across an unordered set of widely separated images by computing view-similarity value between each image pair, so that the WBS matching computation can be limited only among the reasonable view pairs, and the illogical view pairs can be filtered out for later fine matching. A typical method for coarse matching is proposed by Schaffalitzky and Zisserman [10], in which local invariant features of each image are firstly extracted and stored in a features-versus-views hash table. Then a greedy spanning tree for WBS matching is built up according to hash table. Ferrari et al. [1] proposed another algorithm in which the correspondence in two-view is extended to multiple-view and the topological constraint is added to filter out more mismatches. Brown and Lowe at el. [11] presented a system for fully automatic recognition and reconstruction of 3D objects, in which the matches between all unordered images in database are found by using local invariant feature characterized by SIFT descriptor [13]. Further more, Jian Yao et al. [12] presented a more robust view-ordering algorithm in which the views are organized into clusters, then building up the view matching-spanning tree under epipolar constraint, so the further mismatches would be kicked out and result will be more robust.

In this paper, we present an efficient method for the multi-view WBS matching over a large set of unordered images taken from widely separated views. First we propose a maximum a posterior (MAP) model for large-scale 3D reconstruction instead of ML model. Under this MAP model, any available prior knowledge about the unknown parameters can be readily incorporated into the 3D reconstruction process. Secondly, a novel annealing based algorithm for image clustering is developed, by which a large set of unordered images can be efficiently partitioned into a series of clusters of related views (e.g. a part of the whole large-scale scene) by calculating the initial matching numbers of invariant features (found and characterized by SIFT descriptor) between each image pair.

The rest of this paper is organized as follows. The formal deduction and details about MAP model is introduced in section 2. The efficient annealing based ordering algorithm for clustering a large set of unordered images is presented in section 3, and in section 4 we make analyses about the efficiency and computational complexity of the new

ordering algorithm. Finally, the experiment results and comparisons between other ordering algorithms are given in section 5 and the conclusions are drawn in section 6.

2 MAP Model for Large-scale 3D Reconstruction

Most of the problem of computer vision can be treated as structure from motion (SFM) problem [15], especially 3D reconstruction, which means that 3D scene structure and camera motion parameters are recovered and estimated by a series of images taken from different viewpoints.

For this complex geometric parameters estimation problem, we first characterize the parameters set like Dellaert did in [6-7], i.e. we define structure and motion parameters set as Θ, which consists of 3D feature locations X and cameras parameters M; the set of local features measurements as U, and correspondence vector J that records which 2D feature point corresponds to which 3D structure feature point. Suppose that the correspondence is known, the problem can be described by a ML model as follows:

$$\Theta^* = \arg\max \log L(\Theta; U, J) \tag{1}$$

According to Bayesian law, we can employ MAP estimation instead of ML model by incorporating prior knowledge $P(\Theta)$:

$$\Theta^* = \arg\max_{\Theta} \log P(\Theta \mid U, J) = \arg\max_{\Theta} \{\log L(\Theta; U, J) + \log P(\Theta)\} \tag{2}$$

The prior information $P(\Theta)$ in narrow sense can be viewed as the prior probability distribution of parameters; here it can be broadly viewed as any information about 3D reconstruction that is available. In fact, there is little prior knowledge about SFM in most cases and the majority of existing SFM methods assumes no prior knowledge on SFM parameters at all. But in large-scale 3D reconstruction, the whole scene is shot and caught by a large set of wide-baseline images taken from very different viewpoints. For each small part of the huge scene, only part of unordered images is taken for it. One image is low related or irrelevant at all with another one if they are taken in widely-separated views for different part of the huge scene. Therefore, it is unreasonable to make WBS matching between an image pair that is shot for very different parts of the huge scene, and this was the reason that the ML model is not efficient enough and time consuming is huge. As a result, obviously, this fact can be treated as the prior knowledge for large-scale 3D reconstruction.

In order to use the prior information discussed above, we make a partition for the unordered image set, i.e. organize the wide-baseline images into clusters of related sub-scene of the huge scene:

$$G = \left\{ G_i ; G_i \cap G_j = \varnothing, i \neq j \right\} \tag{3}$$

The notation G is the set of all images. After partition, each image subset G_i is mutually exclusive with each other. According to the abovementioned partition, we can make a partition for SFM parameter set as well, and each subset of parameters corresponds to a relative image subset, i.e. we "cut" the huge scene into a number of small scenes. Consequently, Eq.(2) can be further expressed like follows:

$$\Theta^* = \bigcup_{i=1}^{n} \Theta_i = \bigcup_{i=1}^{n} \arg\max_{\Theta_i} \log P(\Theta_i \mid U_i, J_i) = \arg\max_{\Theta} \log(L(\Theta; U, J) P(\Theta)) \tag{4}$$

The typical method to solve this problem is known as Bundle Adjustment, which can be convenient to add kinds of constraints and sparse solver techniques. Noted that the

Eq.(4) is obtained under the assumption that the correspondence is known, when the assumption is incorrect, the formula has to be changed like follows:

$$\Theta^* = \bigcup_{i=1}^{n} \Theta_i = \bigcup_{i=1}^{n} \arg\max_{\Theta_i} \log P(\Theta_i, J_i \mid U_i) = \arg\max_{\Theta} \log(L(\Theta, J; U) P(\Theta)) \qquad (5)$$

In order to perform this MAP without correspondence, a direct way is to marginalize Eq.(5) over all the possible correspondences. But it is impossible to get the posterior distribution $P(\Theta \mid U)$ by doing so, since the direct way would unavoidably suffer from the unacceptable computational cost. Fortunately, Dellaert [6-7] proposed a practical MCEM algorithm for this situation in which the correspondence is unknown. In the MCEM, the unknown correspondence is treated as hidden variable, and the issue of correspondence is solved in parallel with the estimation of the SFM parameters.

According the MAP model discussed above, the large-scale reconstruction can be efficiently cut into a series of small-scale reconstructions, then the MCEM algorithm is employed and tackling the reconstructions in parallel with each other. And finally the high quality 3D model for the whole scene would be obtained by join the reconstructed pieces together.

3 Annealing Clustering Algorithm

In this section, we propose a novel annealing clustering algorithm that can efficiently group a large set of wide-baseline images into clusters. Similar work has been done for small scene reconstruction in [10-13], and the key ideas of them are to compute similarity-values of all image pairs, and to cluster them into different groups and build the spanning tree for matching. However, the computational cost is high when the set of image is large.

3.1 Optimization for Images Set Clustering

Generally speaking, we abstract away from the images clustering problem and think of it in terms of weighted undirected graph. First we define the graph G as $< V, E >$, where the vertices V correspond to the images and the edges E are identified with the relativity between image pairs and the graph is fully connected by the edges E. For each edge $e = (v_i, v_j)$ we define the weight term as follows:

$$w(v_i, v_j) = feature\_matching\_number(v_i, v_j) \qquad (6)$$

In fact, any value that can show how similar or different a image pair are in quantity can be viewed as the weight of edge term. Here we choose the matching number of local invariant features characterized by SIFT descriptor. The more the matching number is, the more similar the two images are.

According to the weight undirected graph defined above, to find a partition C of the images set is equivalent to find a cut set that can cut the graph into several connected sub-graphs, and the images set clustering problem can be conveniently viewed as a optimization problem to minimize the object function, which can be defined as the weight sum of the cut set:

$$W(C) = \sum_{k} w(e_k) = \sum_{i \neq j} w(v_i, v_j) \qquad (7)$$

Considering the convenience for further performing and formulation, we change the object function to be the sum of the weight of the connected sub-graphs, so the minimization becomes maximization instead:

$$W(C) = \sum_{i=1}^{|G|} W(G_i) = \sum_{i=1}^{|G|} \sum_{j=1}^{|G_i|} \sum_{k=j+1}^{|G_i|} w(v_j, v_k) \tag{8}$$

3.2 Annealing Clustering Algorithm

The optimization problem has been well studied for lots of years and there are plenty of existing methods. In this paper, we choose annealing algorithm for the optimization problem mentioned above. The main advantage of annealing algorithm is that it can avoid local extreme and find the global maximum (or minimum) of the object function. The key for efficiency of annealing algorithm is the proposal strategy, which finds out the candidate cluster for choice. In this case, the proposal strategy is related with two factors:

1) The partition for the images set, which decides which images are clustered together.

2) The number of image subset.

Considering these two factors, our strategy and annealing based clustering algorithm can be concluded as follows:

1) Cut the images set into a number of subsets randomly, and each subset has only two images;

2) Calculate the weight $W(C_k)$ for each subset, and the average weight $\overline{W}$.

3) Randomly choose two images subset G_i and G_j whose weights $W(G_i)$ and $W(G_j)$ are lower than the average weight $\overline{W}$, choose a image in each subset at random and swap them, repeat this step for T_i times.

4) Accept the new clustering $C' = C_{k+1}$ with the following probability:

$$\exp\left[-\frac{(W(C') - W(C_k))}{T}\right] \tag{9}$$

5) Unite two subsets if the images of them are high related. Repeat the step until the number of subset is N_i.

During the iteration of annealing, we first increase the cluster number N_i and the images swapping number T_i, and then these two parameters would gradually decrease with the temperature parameters. In the internal iteration, images are swapped within clusters to increase their weights so that each cluster can be as cohesive as possible. In the external iteration, clusters are trying to be combined if images within them are highly related. Both these steps increase the object function in intuition and the maximum (or minimum) can be eventually found.

4 Experimental Result and Comparison

In order to verify the proposed grouping algorithm, we have made a series of experiments on various image data comes from different databases. The result shows that our approach works as efficiently as expected. In this Section, we will show three

representative experiments of them and make comparison with traditional exhaustive coarse matching methods.

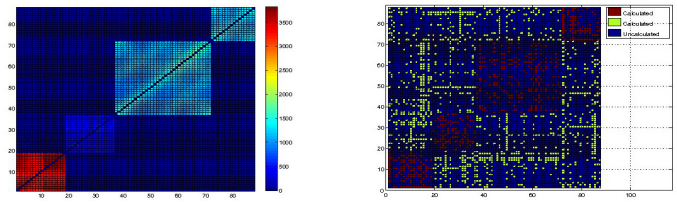
4.1 Unordered Image set with different Rotation

The first experiment is performed on 88 photos which are taken from in 4 dissimilar scenes with different rotation. All of them come from Mikolajczyk's database [15]. The examples of them are shown in Figure 1. The number is the index for each photo in group. Although the photos are ordered by indexes for convenient to show， the initial input sequence is randomly organised and the algorithm dose not use the index information, therefore, the photo group is still unordered.

Figure 1: Example of each photo group.

In order to validate the efficiency of our new method, we use a view-vs-view table proposed in [10] to make a comparison. The table can directly show the number of initial two-view connections found between the views of the image set. In fact, the table can be seemed as an upper-triangular matrix. Here we change it into a symmetric matrix for programming convenience.

Another special table we define is "cache matrix". It and symmetric matrix is the same type matrix that used to record the initial matching number between image pairs. During the process of the grouping, the program will first try to access the cache matrix when a coarse matching is needed, and the initial matching number will be stored into the matrix if it has not been computed before. Therefore, when the grouping process is over, the matrix can be used to show computation cost that our new algorithm takes for photos grouping.



(a) (b)

Figure 2: 88x88 symmetric matrix (a) and relative cache matrix (b). Each element stands for the initial number of specific two-image connections. In (b), the elements in red and green color mean the corresponding connection numbers have been computed during the grouping process; on the contrary, the blue ones haven't been computed.

Both matrixes of this experiment are shown in Figure 2. For traditional exhaustive matching approaches in [10-13], the symmetric matrix (Fig 2(a)) must be fully computed; for our new approach, only small part of it needs to be computed. By scanning the cache matrix in Fig 2(b) and counting the unrecorded elements, there are

80.17% elements have not been computed, in other words, our grouping algorithm can save more than eighty percent of computational cost comparing with old exhaustive ones.

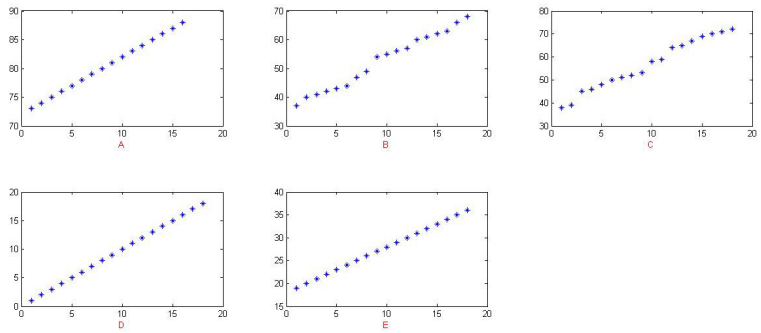


Figure 3: Grouping result. The 88 photos are grouping into 5 clusters. And photos in cluster B and cluster C are actually taken from the same scene.

In this experiment, the extern-iteration number of grouping algorithm is set to 2 and the inner-iteration number is 200. The grouping result is shown in Figure 3. Careful reader may find that the photo grouping result for "New York" is divided into two subgroups B and C. It is not surprising, because we only perform 2 external iteration times, for the grouping algorithm, which is actually an annealing one, is too short and "hot" for it to find a global optimal sorting result. But considering about the time cost, this local optimal result we get is acceptable because it did not sort the different photos into one group.

4.2 Unordered Image set of Large-Scale 3D Scene

The second experiment is performed on 46 photos that also come from [15]. All the photos are taken from different viewpoints in 4 dissimilar scenes. The examples of them are shown in Figure 4 and the symmetric matrixes are shown in Figure 5.

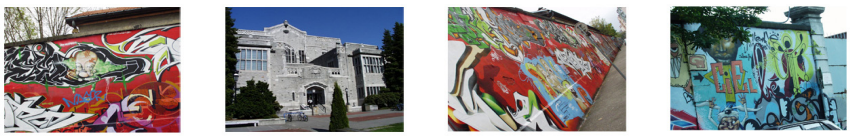
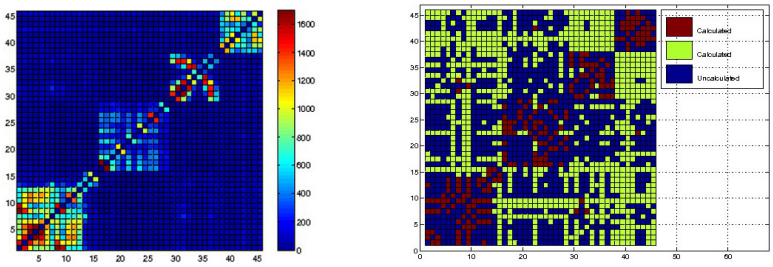


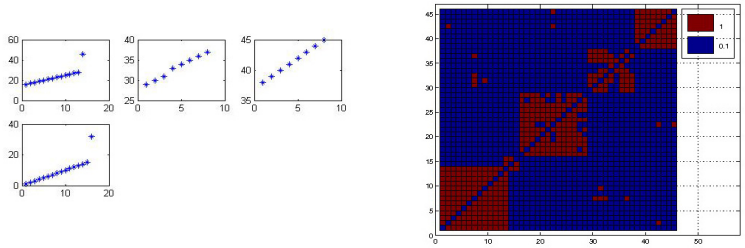
Figure 4: Examples of each photo group

The grouping result is shown in Figure 5, which is fairly perfect except a few mistakes. It is easy to explain that, since the initial matching number is used to be the relativity measure between image pair, and sometimes these initial numbers between irrelevant photos are even larger than the relative ones. This flaw is more evident if the symmetric matrix is binarized, as shown in Figure 6(b). For example, the 7th photo which belongs to the "Graff4" group has the unreasonable high similarity with the 30th to 32nd photos in "Graff5" group, as shown in Figure 5(b), which finally led to the grouping error (the isolated point shown in Figure 6(a)).



(a) (b)

Figure 5: 46x46 symmetric matrix (a) and relative cache matrix (b). By counting the blue elements in (b), we can conclude that the computation cost reduces by 51.50%.



(a) (b)

Figure 6: (a) Grouping result (b) binarized result of symmetric matrix in Fig 5 (a).

4.3 Unordered Images of Large-Scale 3D Scene

The third experiment is performed on the real image set taken by authors in Xi'an city, China, as shown in Figure 7, (the scene can be seen by Google Earth at 34°14′10.17″N, 108°54′05.72″E). Instead of showing matrixes and graphs, we show the grouping result of real images clusters in Figure 8.

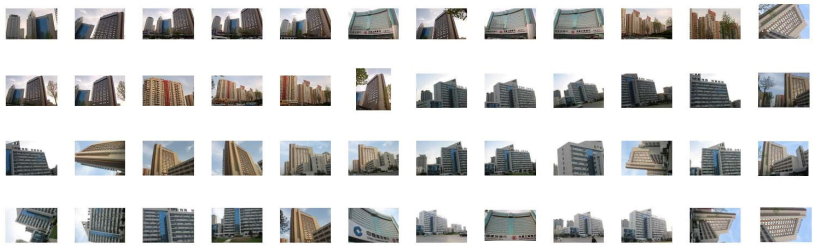


Figure 7: Real images taken by authors. There are total 48 photos for this city urban scene, which can be cut into 5 pieces of building communities. Note: these photos are organized into an unordered set deliberately.

In this experiment, the 48 unordered photos are taken from widely different viewpoints and have large affine warping. Therefore, the grouping process iterates for more than 50 times and saves only 24.02% computation cost in total. It eventually groups them into 5 clusters. However, since there are a lot of repeat structures leading

to wrong feature matching, there are 4 photos grouped into wrong clusters (marked by red frame).

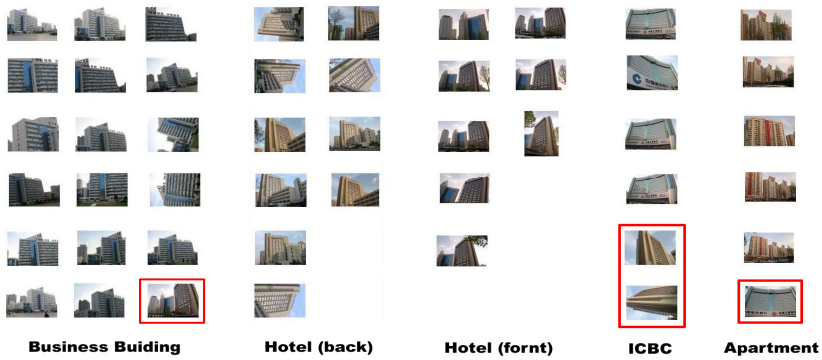


Figure 8: Photos clustering result.

5 Conclusion

In this paper we have presented a novel MAP model which can easily incorporate prior knowledge of images clustering and an annealing based clustering algorithm which can efficiently organize large number of images into clusters. The model is efficient and suitable for large-scale 3D reconstruction. Moreover, the MAP model can be viewed as a general framework and the specific form can be changed when using different camera, re-projection and noise models.

For the annealing clustering algorithm, experimental results show that the algorithm is efficient to organize a large set of unordered images and is convenient to implement. In this paper, we use the matching number of invariant SIFT features. In fact, any value can be used if it can express the view-similarity of image in quantity. As a result, the MAP model along with the clustering algorithm can be used to model any SFM parameters estimation problem if the parameters can be "grouped".

Acknowledgments

This work is supported by National Hi-Tech Development Programs of China under grant No. 2007AA01Z314, National Natural Science Fund (60403008) and Program for New Century Excellent Talents in University (NCET-06-0882), P. R. China.

References

[1] V. Ferrari, T. Tuytelaars, L.J. Van Gool, Wide-baseline multiple-view correspondences, IEEE Conference on Computer Vision and Pattern Recognition (CVPR03), vol. 1, Wisconsin, Madison, USA, June 2003, pp. 718–725.

[2] D. Tell, S. Carlsson, Wide baseline point matching using affine invariants computed from intensity profiles, Europe Conference on Computer Vision (ECCV00), vol. 1, Trinity College Dublin, Ireland, June 2000, pp. 814–828.

[3] P. Smith, D. Sinclair, R. Cipolla, K. Wood, Effective corner matching, British Machine Vision Conference (BMVC98), vol. 2, Southampton, UK, September 1998, pp. 545–556.

[4] X. Lu, R. Manduchi, Wide-baseline feature matching using the cross-epipolar ordering constraint, IEEE Conference on Computer Vision and Pattern Recognition (CVPR04), vol. 1, Washington DC, USA, July 2004, pp. 16–23.

[5] N. Snavely, S. Seitz, R. Szeliski, Photo tourism: Exploring Photo collections in 3D. ACM Transactions on Graphics (SIGGRAPH Proceedings), 25(3):835--846, 2006.

[6] F. Dellaert, Monte Carlo EM for Data-Association and its Application in Computer Vision. PhD thesis, Carnegie Mellon University, 2001.

[7] F. Dellaert, S. Seitz, C. Thorpe, and S. Thrun, Structure from motion without correspondence. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR00), Vol. 2, Hilton Head, Island, June, 2000, pp. 557-564.

[8] S. Zhu, F. Dellaert, Z. Tu, Markov Chain Monte Carlo for Computer Vision - A tutorial, IEEE Conference on Computer Vision (ICCV05), 2005.

[9] F. Dellaert, Out-of-Core Bundle Adjustment for Large-Scale 3D Reconstruction, IEEE Conference of Computer Vision (ICCV07), 2007, pp.1-8.

[10] F. Schaffalitzky, and A. Zisserman, Multi-view Matching for Unordered Image Sets, or "How Do I Organize My Holiday Snaps?", IEEE Europe Conference on Computer Vision (ECCV02) vol. 1, Copenhagen, May 2002, pp. 414-431.

[11] M. Brown, D.G. Lowe, Unsupervised 3D object recognition and reconstruction in unordered datasets, Int Conf. on 3-D Digital Imaging and Modelling, Canada, 2005, pp. 56-63.

[12] Jian Yao, Wai-Kuen Cham, Robust multi-view feature matching from multiple unordered views, Pattern Recognition, Vol.40, No.11, 2007, pp.3081-3099.

[13] D.G. Lowe, Distinctive image features from scale-invariant key points, Int. J. Computer Vision Vol.60, No.2, 2004, pp. 91-110.

[14] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2000.

[15] http://lear.inrialpes.fr/people/Mikolajczyk/