STRUCTURE FROM MOTION USING UNORDERED SETS OF IMAGES

Ruxandra ȚAPU[1], Bogdan MOCANU[2], Ermina ȚAPU[3], Teodor PETRESCU[4]

In this paper we propose a novel method for interest point extraction and matching with high confidence scores in the context of 3D object reconstruction from multiple images taken from the same video camera. We start by using pyramidal FAST algorithm to detect image features that are further described using SIFT method. Then, we determine high confident matching by employing the RANSAC technique. Finally we propose a recursive algorithm that extends the set of inliers using local homographies. Our framework is able to handle important camera movement, object occlusions and image noise. The experimental evaluation performed on various challenging image sets shows significant improvements of the SfM when applying the proposed strategy.

Keywords: Structure from motion, interest point matching, RANSAC algorithm, local homography estimation

1. Introduction

Three-dimensional representation and reconstruction of real life objects starting from multiple views has been an active topic of research in the area of artificial intelligence. The process, also known as Structure from Motion (SfM) is based on photogrammetric principles. So, by using a set of images, taken by an uncalibrated camera that represents different perspective of the same rigid object or scene, the objective is to automatically recover the 3D structure of the environment [1]. Increasing demands from the virtual reality, navigation, robotics, medical and film production industries have resulted in major developments over the last twenty years.

In this paper we tackle the issue of SfM focusing our attention on feature point matching, represented by interest points selected from each individual

[1] Lecturer, Faculty of Electronics, Telecommunications and Information Technology, University POLITEHNICA of Bucharest, Romania, e-mail: ruxandra_tapu@comm.pub.ro
[2] Lecturer, Faculty of Electronics, Telecommunications and Information Technology, University POLITEHNICA of Bucharest, Romania, e-mail: bcmocanu@comm.pub.ro
[3] Associate Professor, Faculty of Aerospace Engineering, University POLITEHNICA of Bucharest, Romania, e-mail: ermina.tapu@upb.ro
[4] Professor, Faculty of Electronics, Telecommunications and Information Technology, University POLITEHNICA of Bucharest, Romania, e-mail: teodor.petrescu@electronica.pub.ro.

image, which can severely alter the reconstructed objects especially for large reconstruction of 3D scenes. Thus, we propose a novel method to address it. In the presence of occlusion, image noise, fast camera movements, motion blur or object leaving the camera's field of view, occasional feature mismatching or dropouts may appear. These problems make developing a robust feature matcher system very challenging. To our very best knowledge the impact of inconsistent feature matching in the framework of SfM has not been thoroughly studied in the technical literature.

The rest of the paper is organized as follows: Section II presents a short review of the technical literature dedicated to 3D reconstruction. In Section III we introduce a novel framework for structure from motion, dedicated to rigid objects, using unordered image sets. Section IV presents the experimental results obtained on various datasets publically available in the state of the art literature. Finally, Section V concludes the paper and provides perspectives for further development.

2. State of the art review

Various structure from motion algorithms were proposed in the technical literature trying to estimate the photographic camera parameters and to develop a sparse 3D representation of the scene geometry by using ordered/unordered image sets. The first step of the framework is images overlapping and correspondences matching across different perspective of the scene. In the second step, by using the correspondences between interest points the view are geometrically related. Finally, base on the epipolar geometry the camera parameters and the 3D scene structure can be estimated with some degree of error.

The problem of image correspondence was intensely studied on the last couple of years leading to the development of sequential matching algorithms. Some approaches as [2] and [3] detect local features and match them using local descriptors: SIFT (Scale Invariant Feature Transform) [4] or SURF (Speeded Up Robust Features) [5]. Different authors in [6] and [7] propose using tracking algorithms as LK (Lucas-Kanade) [8] in order to create small baseline triangulation.

The algorithms based on invariant features and Lucas Kanade tracker are sensitive to occlusions, reflection, zoom level, etc. Generally, sequential matchers are not robust to important camera movement, which translates into large image transformation. The problem becomes particularly difficult in the presence of repetitive elements that give rise to multiple and ambiguous correspondence. Unfortunately, such correspondence between image matches makes the scene structure estimation an unstable process and potentially will lead to poor reconstruction results. In [9] by using the graph-connectivity across huge image collection, Hartley identifies links between image pairs viewing the same or

similar objects. In [10] the authors propose to use typically observed redundancy and implement a graph structure to encode visual relation in images. By chaining the (reversible) transformations over cycles in this graph, they build a suitable statistics to identifying inconsistent loops and infer false matches. However, the authors are not consistent in treating the problem of repetitive structures, which can severely influence the quality of the reconstructed scene.

In order to overcome the above limitation different authors [11] use ASIFT [12] descriptors to improve the feature matching performance under substantial viewpoint change. In [13] Engels proposed integrating wide-based local features to improve the Structure from Motion (SfM). The method is able to correctly create small and independent submaps but only for a reduced number or images. In a large dataset the method cannot produce long and accurate point tracks. In comparison, the method proposed in this paper can effective develop high-quality point track estimation.

3. Proposed approach

We start our framework by using the pyramidal FAST algorithm [14] in order to detect interest features in a given set of N images (I) which represent different views of the same object. The interest points are further described using the SIFT algorithm firstly introduced in [4]. The features from one image I_n are matched against all features extracted from the set of images. One of the advantages of working with representative points consists on the system obliviousness to the scene content (*i.e.* the scene can have any structure with any texture as long as the motion is a single rigid body).

Given f_n an interest point in image I_n and its associated descriptor $d(f_n)$, we used a two-nearest-neighbor search strategy in order to determine if in another image I_{n+1} exists a corresponding similar point f_{n+1} characterized by its associated descriptor $d(f_{n+1})$. So, we establish the *2NN* features for f_n in image I_{n+1} by using the *L2* norm distance between descriptor vectors. We denote them with $N_1^{n+1}(f_n)$ and $N_2^{n+1}(f_n)$. Then, we compute a matching confidence score (c) in order to establish the global distinctiveness between correspondent features:

$$c = \frac{\left\| d\left(N_1^{n+1}(f_n) - d(f_n)\right)\right\|}{\left\| d\left(N_2^{n+1}(f_n) - d(f_n)\right)\right\|} \tag{1}$$

If $c < Th_1$ we consider the match of f_n with $f_{n+1} = N_1^{n+1}(f_n)$ as correct (Fig. 1a). We set in our experiments $Th_1 = 0.7$. The computational complexity of this step is $O(Nr_n \cdot Nr_{n+1})$, where Nr_n is the number of features in image I_n. However, when repetitive structures are presented in the scene or images are distorted by noise it becomes difficult to find correct matches even in very similar images.

Now, we propose to determine high confidence matches between images (I_n and I_{n+1}) and remove outlier points. We used the RANSAC algorithm [15] to determine the fundamental matrix $F_{n,n+1}$ that estimates the geometrical transformation parameters from one image to another. We randomly select a minimal set of features from the entire set of correspondences, estimate a transformation and then determine how well the computed matrix works for the entire set of matches (Fig. 1b). Interest points satisfying the transformation are labeled as inliers (Ω). The minimal set of points considered is eight, which proves to be more stable in the presence of noise.

However, if significant image distortion exists, the above strategy significantly reduces the number of inliers that translates into a low quality of the reconstructed 3D object. To address this problem, we introduce next a reinsertion method that robustly identifies missed matches.

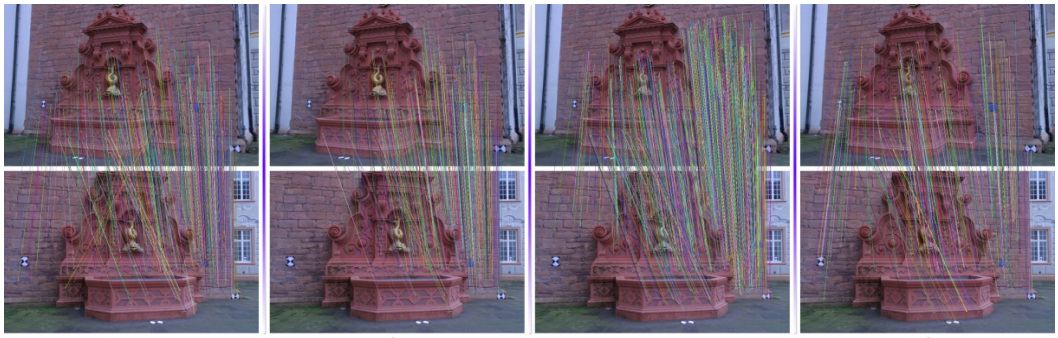

Fig.1. Matching strategy comparison. (a). Interest points resulted when applying the global distinctiveness constraint. 11183 features are extracted using FAST detector, but after this step only 454 matches are retained; (b). Matches between interest points that satisfy the geometrical transformation (340); (c). Final results obtained after applying our algorithm (1690 matches are conserved); (d). Results obtained using the ASIFT descriptor (880 matches are retained).

Because many interest points can present similar motions we extend the set of inliers using local homographies $\{H_k^{n,n+1} | k = 1, \dots, M\}$ estimated on image regions. We introduce a recursive algorithm that uses as inputs the set of inliers Ω. Then, we apply the RANSAC algorithm to estimate the homography $H_k^{n,n+1}$ that determines the maximum number of inliers (Ω_1). A new set $\{\Omega_2\} = \{\Omega\} \setminus \{\Omega_1\}$ is

obtained by removing the points satisfying the transformation from Ω. If the size of $\Omega_2 < Th_2$ the process stops, otherwise is repeated recursively with Ω_2 as a new input.

Then, we consider the initial set of features (f_n) detected in image I_n using the pyramidal FAST algorithm and we rectify them with every local homography matrix $H_k^{n,n+1}$ in order to estimate their position in image I_{n+1}.

$$p_{f_{n+1}}^{est} = \left\{ H_k^{n,n+1} \cdot p_{f_n} \middle| k = 1, \dots, M \right\} \tag{2}$$

where $p_{f_n}[x_{f_n}, y_{f_n}, 1]^T$ is the point position expressed in homogenous coordinates and M is the total number of homographic matrices obtained between two image pairs.

The matching error is defined as the difference between the estimated location of the interest points and the actual position determined using the brute force matching strategy:

$$err\left(p_{f_{n+1}}^{est}, p_{f_{n+1}}\right) = \left\| p_{f_{n+1}}^{est} - p_{f_{n+1}} \right\| \tag{3}$$

If $err\left(p_{f_{n+1}}^{est}, p_{f_{n+1}}\right) < Th_3$, the corresponding point is reinserted into the inliers set. Incorrect homographies are unlikely to return high confidence matches. In our experiments we set Th_3 value to 1.5 pixels. In Fig. 1c we present the experimental results obtained after applying our strategy. Fig. 1d gives the matching results of the ASIFT [12] descriptor.

After reinserting all correct matches we estimated next the camera poses.

In the following section, we present our method to estimate the interest points positions in the 3D space starting from 2D matches obtained between any image pairs from the dataset. Because, the 3D estimation starting from 2D is not an invertible process the task of 3D reconstruction from images is very challenging. In our development we have considered the perspective rays converging into the camera center which translated into 3D point projection on an image plane.

The pinhole camera model establishes a relation between the 3D point and its correspondence on the 2D image. First, the rigid body transformation is computed. This relates the 3D point (Pt_{3D}) expressed in homogenous coordinates $Pt_{3D} \sim [X\ Y\ Z\ 1]^T$ to the point Pt_{CC} expressed in the camera coordinate system $Pt_{CC} \sim [X_C\ Y_C\ Z_C\ 1]^T$:

$$\begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix} \sim \begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{4}$$

where **T** is a tridimensional vector representing the camera translation, **R** is a 3 x 3 rotation matrix that giving the camera orientation, while ∼ defines an equality up to scale.

The correspondence between $Pt_{CC} \sim [X_C\ Y_C\ Z_C\ 1]^T$ and the 2D point ($\mathrm{Pt_{2D}} \sim [x\ y\ 1]^T$) on the camera image plane is determined based on the 3D to 2D transformation:

$$x = f\frac{X_c}{Z_c}\ ; y = f\frac{Y_c}{Z_c}\ , \tag{5}$$

where *f* is the camera focal length. The focal length can be considered in direct correlation with the scale factor encountered in the camera calibration process. Then Eq.4 can express as:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix}, \tag{6}$$

Finally a 2D to 2D transform is performed that relates points in the camera plane to pixel coordinates $\mathrm{Pt_{2DP}} \sim [u\ v\ 1]^T$ as follow:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \boldsymbol{K} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha_u & s & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{7}$$

where α_u and α_v are the scale factors, $\mathrm{Pt_{2D0}} = [u_0\ \ v_0\ ]^T$ is the principal point and *s* is the skew. Equation (4)-(7) can be combined into a single linear equation:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} \alpha_u & s & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \boldsymbol{P} \cdot Pt_{3D}\ , \tag{8}$$

where **P** is a 3 x 4 projection matrix.

Now, by knowing the projection of a 3D point into an image, its projection into a second image is restricted to the epipolar line.

So, given a Pt'_{CC} in the coordinates system of a camera C' its position Pt_{CC} in the coordinate system of the camera C can be computed as:

$$Pt_{CC} = \boldsymbol{R} \cdot Pt'_{CC} + \boldsymbol{T} \leftrightarrow 0 = Pt_{\boldsymbol{CC}}^{\boldsymbol{T}} \cdot [\boldsymbol{T}]_X \cdot \boldsymbol{R} \cdot Pt'_{CC} = Pt_{\boldsymbol{CC}}^{\boldsymbol{T}} \cdot \boldsymbol{E} \cdot Pt'_{CC} = 0, \quad (9)$$

where $\boldsymbol{E} \sim [\boldsymbol{T}]_X \cdot \boldsymbol{R}$ is a 3 x 3 essential matrix and $[\boldsymbol{T}]_X$ is the cross product matrix. Equation (9) holds also for image points, giving the epipolar constraint.

From equation (7) image points $\mathrm{Pt_{2D}}$ can determine the pixels position $\mathrm{Pt_{2DP}}$ by using inverse camera calibration matrix $\mathrm{Pt_{2DP}} \sim \mathrm{K}^{-1} \cdot \mathrm{Pt_{2D}}$. Applying this observation to the epipolar constrains results that:

$$(K^{-1} \cdot \mathrm{Pt_{2DP}})^T \cdot \boldsymbol{E} \cdot (K'^{-1} \cdot \mathrm{Pt'_{2DP}})^T = 0 \leftrightarrow \mathrm{Pt}_{2\mathrm{DP}}^T \cdot \boldsymbol{F} \cdot \mathrm{Pt}_{2\mathrm{DP}}'^T = 0, \quad (10)$$

where $\boldsymbol{F} \sim K^{-1T} \cdot \boldsymbol{E} \cdot K'^{-1}$ is a 3 x 3 matrix of rank 2 entitled fundamental matrix. Then, by knowing the camera calibration matrixes $\boldsymbol{K}$ and its inverse $\boldsymbol{K'}$ we can recover from **F** the essential matrix **E**:

$$\boldsymbol{E} = \boldsymbol{K'^T} \cdot \boldsymbol{F} \cdot \boldsymbol{K} \quad (11)$$

Using the Singular Value Decomposition algorithm, matrix $\boldsymbol{E}$ can be decomposed into a skew symmetric matrix corresponding to translation and an orthonormal matrix corresponding to rotation between views.

Finally, the triangularized vertices are used to construct a depth map which is iteratively refined with bundle adjustment [16].

4. Proposed approach evaluation

We tested the proposed SfM methodology on various real world data acquired in indoor and outdoor scenarios. Fig. 2 presents the results obtained on benchmark images from [17] with varying complexity of the underlying symmetries. The image collections, called Fountain-P11, Leuven castle –LC9, Herzjesu – H8 and Medusa – M19 include 11, 9, 8 and 19 images, respectively. For each sequence, the output model includes the computed camera positions as well as the set of observed 3D points.

In Table 1 we give a complete evaluation of the proposed method, on the considered dataset, in terms of mean reprojection error (MRE) and N3D (number

of 3D points), with respect to the ASIFT descriptor.

Table 1

Number of 3D points and mean reprojection error (pixels) of the proposed method compared with the ASIFT descriptor

Model name	No of images	ASIFT		Proposed method	
		MRE	N3D	MRE	N3D
Fountain	11	1.576	21480	1.247	27528
Leuven castle	9	0.981	28452	0.924	30824
Herzjesu	8	0.813	15205	0.786	15415
Medusa	19	1.572	37184	1.518	58864
Venus de Milo	71	4.315	215013	3.543	245278
Duomo in Pisa	56	2.984	123856	2.776	143586
Notre Dame de Paris	55	2.885	103213	2.432	143870
Temple	78	3.432	278217	3.183	310284
Dino	37	2.471	89431	2.511	120843

Let us analyze the reconstructed models illustrated in Fig. 2. The camera is moving around different objects existent in the real life, in order to capture all perspectives of models. *Note:* it is important to mention that our method is designed for static objects of interest. Other moving objects existent in the scene will not influence the overall performance of the system.

We present the experimental results obtained in two independent situations when extracting and matching feature points based on: (a) the ASIFT descriptor [12] and (b) the strategy introduced in this paper using the pyramidal FAST detector. In order to offer a qualitative evaluation of the proposed methodology we analyze the system performance in terms of mean reprojection error (MRE expressed in pixels) and the total number of 3D points (N3D) of the SfM model. In both cases we assume the set of image unordered.

As is can be observed, from Fig. 2, our system returns a MRE of 1.247 pixels for Fountain-P11 and 0.786 pixels for Herzjesu – H8, while for ASIFT the MRE is 1.576 and 0.813 pixels for the considered set of images. Regarding the number of reconstructed 3D points our system increase the SfM quality with more than 6000 points for the Fountain-P11 and 21000 for Medusa – M19, respectively.

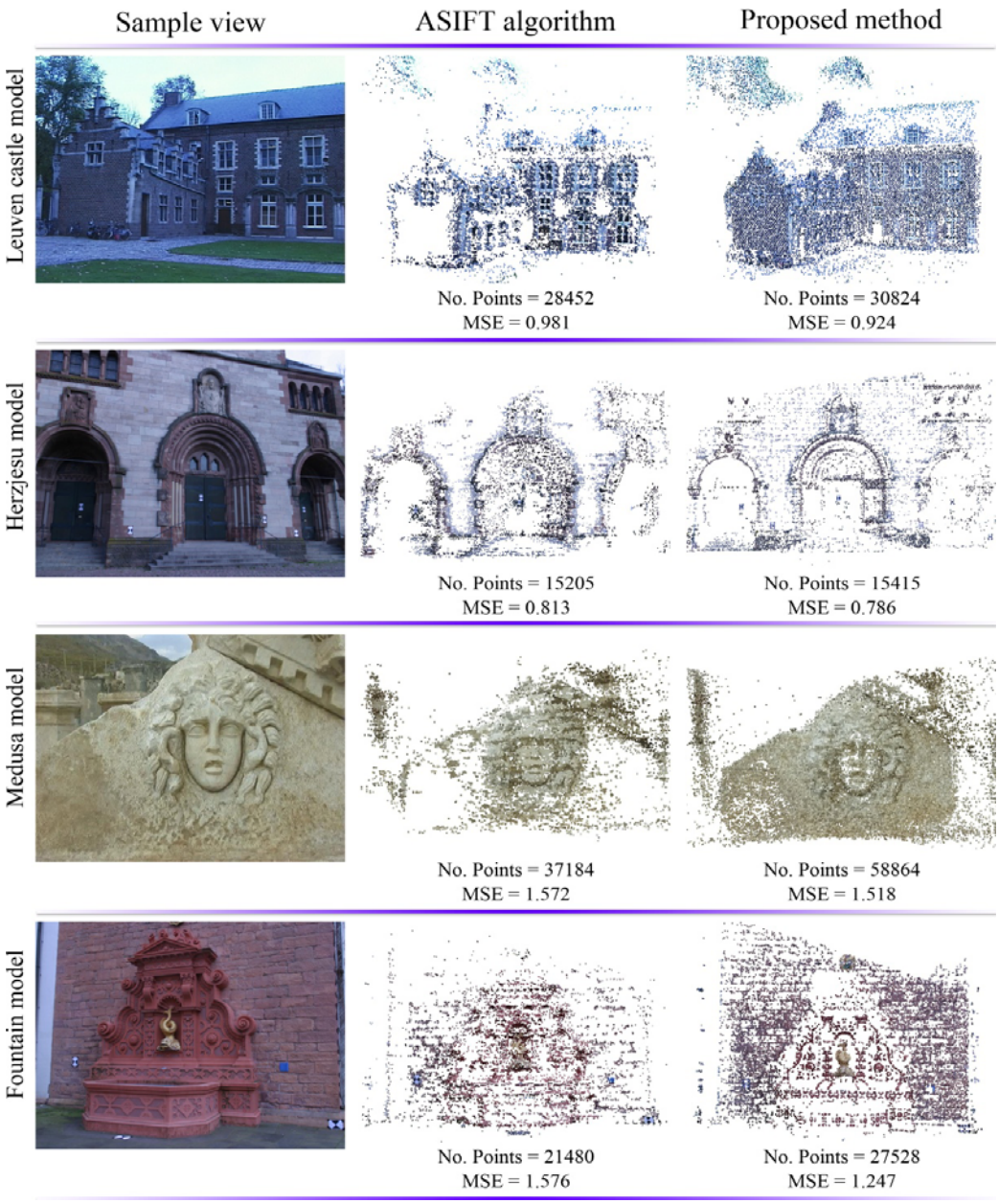

Fig. 2. Structure from motion experimental results comparison when using our algorithm and ASIFT feature descriptor.

The system is implemented in C++ and run on an Intel Xeon Machine 3.6 GHz machine, 16 GB RAM with NVIDIA Quadro 4000 video board under Windows 7. The average processing time for 11 images with the resolution 768 x 576 pixels is around 28 seconds.

5. Conclusions and perspectives

In this paper we propose a complete framework for 3D reconstruction of rigid scenes and objects. The system introduces a novel algorithm for robust estimation and matching of interest points, extracted using FAST and further described using SIFT methods, between image pairs. The proposed technique conserved the local homographs between images and develops a strategy for interest point reinsertion based on high confidences. Different from the classical matcher (LK) we introduce a method of establishing invariant features that allows reducing the matching sensitivity to noise and image distortion.

The experimental evaluation performed on various outdoor scenarios demonstrates the improvement brought by our system in the context of SfM applications for middle and large scale reconstruction of 3D scenes and objects.

For future work we consider extending the proposed system to handle non-rigid (deformable) or dynamic objects and to transfer it on a smartphone device.

Acknowledgment

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/132395 (InnoRESEARCH).

REFERENCES

[1]. *A. Irschara, C. Zach, J. M. Frahm, H. Bischof*, "From structure-from-motion point clouds to fast location recognition", In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[2]. *S. Lazebnik, C. Schmid, J. A. Ponce*, "A sparse texture representation using local affine regions", IEEE Transaction on Pattern Analysis and Machine Intelligence, **vol. 27**, pp. 1265-1278, 2005.

[3]. *J. Matas, O. Chum, M. Urban, T. Pajdla*, "Robust wide-baseline stereo from maximally stable extremal regions", Image Vision Computing, **vol. 22**, pp. 761-767, 2004.

[4]. *D. Lowe*, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, pp. 1-28, 2004.

[5]. *H. Bay, T. Tuytelaars, L. Van Gool*, "SURF: speeded up robust features", In Proc. IEEE European Conference on Computer Vision (ECCV), 2006.

[6]. *J. Shi, C. Tomasi*, "Good features to track", In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 593-600, 1994.

[7]. *C. Zach, D. Gallup, F. Michae-Frahm*, "Fast gain-adaptive klt tracking on the gpu", In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Visual Computer Vision on GPU's, 2008.

[8]. *S. Baker and I. Matthews*, "Lucas-Kanade 20 Years on: A Unifying Framework: Part 1," Int'l Journal of Computer Vision, **vol. 56**, no. 3, pp. 221- 255, 2004.

[9]. *K. Hartley, N. Gelfand, O. Vsjanikov, M. Anjaneya, L. Guibas*, " Image webs: Computing and exploiting connectivity in image collections", In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3432 –3439, 2010.

[10]. *C. Zack, M. Klopschitz, M. Pollefeys*, "Disambiguating visual relations using loop constraints", In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1426 –1433, 2010.

[11]. *A. Irschara, C. Zach, J.M. Frahm, H. Bischof*, "From structure-from-motion point clouds to fast location recognition", In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[12]. *J.M. Morel and G. Yu, "ASIFT*: A New Framework for Fully Affine Invariant Image Comparison" SIAM Journal on Imaging Sciences, 2009.

[13]. *C. Engels, F. Fraundorfer, D. Nister*, "Integration of tracked and recognized features for locally and globally robust structure from motion", In VISAPP (Workshop on Robot Perception), pp. 13-22, 2008.

[14]. *E. Rosten and T. Drummond,* "Machine learning for high speed corner detection", in In Proc. IEEE European Conference on Computer Vision, **vol.1**, 2006.

[15]. *J. Lee and G. Y. Kim,* "Robust estimation of camera homography using fuzzy RANSAC", ICCSIA, 2007.

[16]. *B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon*, "Bundle Adjustment-A Modern Synthesis", *LNCS*, **vol. 1883**, pp. 298-375, 2000.

[17]. *C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen*, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.