Automatic Gigapixel Mosaicing in Large Scale Unstructured Underground Environments

Daniela Craciun[1,2], Nicolas Paparoditis[1], and Francis Schmitt[2]
[1]Institut Géographique National, 2/4 avenue Pasteur, F-94165 Saint Mandé, France
firstname.lastname@ign.fr
[2]Télécom ParisTech, CNRS URA 820-TSI Dept, 46 rue Barrault, F-75013 Paris, France
firstname.lastname@telecom-paristech.fr

Abstract

We are currently developing a vision-based system aiming at automatically generating in situ ultra-high resolution mosaics in previously unknown, complex and unstructured underground environments. Nowadays, image stitching algorithms present several limitations when dealing with unstructured environments. The most important to our concern is the ability to deal with feature-less areas. In this paper we describe an automatic on line Gigapixel mosaicing system capable to deal with the absence of reliably detectable and trackable features. The input of our algorithm is a pose-annotated sequence of high-resolution images acquired from a common optical center by a calibrated pan-tilt motorized digital camera unit. The proposed mosaicing system is powered by a global-to-local pairwise image alignment which recovers the rotations relating the overlapping images in a coarse-to-fine approach. The local motion procedure outputs a list of locally matched anonymous features which are later injected in a bundle adjustment engine for multi-view fine alignment. The proposed algorithm combines the state of the art mosaicing techniques in a complementary and efficient fashion providing an environment-independent solution for the image mosaicing task. The final output of the algorithm is a Gigapixel spherical mosaic rendering. Tests on real data acquired in a prehistorical cave (Tautavel Cave, France) illustrate mosaicing examples obtained from several hundreds of high-resolution images.

1 Introduction and Motivation

The research work presented in this paper addresses the multi-view image alignment problem for automatic *in situ* generation of spherical Gigapixel mosaics in complex and previously unknown unstructured underground environments. This work is motivated by a vision-based system under development aimed at generating *in situ* photorealistic 3D models of high-risk and "difficult-to-access" areas without requiring human operator intervention.

In challenging environments several needs must be fulfilled in order to improve the capabilities of the nowadays image mosaicing algorithms. In this paper we address key issues for automatic mosaicing in unstructured environments, such as: dealing with the absence of reliably detectable and trackable features, handling noisy initial guess provided by the physical instrumentation, and robustness to occlusion, illumination changes and blur. Special attention must be also given to both constraints, time and in-situ access, therefore assuring complete scene mosaicing and real-time performances are majors concerns.

Our paper is structured as follows. Section 2 provides a concise overview of prior work, followed by the description of the proposed image mosaicing system in Section 3. The next section presents experimental results, summing up with conclusion and future work in Section 5.

2 Related Work

Image stitching was pioneered back in 1970s. Since then, the image mosaicing theory has been intensively addressed and considerably improved by researchers [9, 1, 4, 3] and commercial groups [2, 8]. A recent survey of the existent image mosaicing techniques can be found in [9]. The fundamental ingredient in mosaic computation is the image alignment process which consists in computing the 3D Euclidian rigid transformation $\mathbf{T}$ which lies between overlapping images. Typical methods minimize either radiometric or geometric error over the overlapping area. Two main approaches are usually employed for the image alignment task. Direct approaches [4] are accurate but unaffordable for high resolution images even if a close initial estimation of $\mathbf{T}$ is given. If no initial estimation of $\mathbf{T}$ is given, feature-based methods [1] are preferred. These methods rely on feature extraction and matching followed by an outliers rejection step via the RANSAC [5] procedure. Feature-based techniques are instable in regions that are either too homogeneous (such as sky portions) or too textured. Thus, the algorithm fails to match images that should be aligned or to fit an accurate and outlier-free image motion model. While improving the state of the art these methods remain limited to the application field: image mosaicing in urban structured scenes.

3 Gigapixel Mosaicing System

The inputs of our algorithm are several hundreds of ordered high resolution images acquired from a common optical center. The capturing device illustrated in figure 1 is previously parameterized with the field of view to be cover and the desired overlap between adjacent images. The method proposed in this paper uses the complementarity of the existing image alignment techniques (direct vs. feature-based) and fuses their main advantages in an efficient fashion.

(a) (b)

Figure 1: Acquisition System: a NIKON D70 digital camera (a) with its optical center fixed on a motorized pan-tilt head attached to a tripod base (b).

First, a global-to-local pairwise motion estimation is performed which refines the initial estimates provided by the pan-tilt head. We solve for rotation using a pyramidal patch-based correlation procedure via quaternions.

In order to provide robustness to deviations from pure parallax-free motion[1], the global rotation initializes a patch-based local motion estimation. The pairwise procedure outputs a list of locally matched image points via a translational motion model. Since the matched points do not correspond to any corner-like features, we introduce them as *anonymous features* (AF).

Second, the multi-view fine alignment is achieved by injecting the AF matches in a bundle adjustment engine [10].

The proposed scheme detains several advantages over the existing methods. Comparing to Teller's approach [4], our method can handle very noisy initial guess and big amounts of parallax. Moreover, the pyramidal patch-based framework enables fast high resolution image matching which is a key aspect for the Gigapixel mosaicing task. In addition, the bundle adjustment scheme enables final optimization with real-time performances. Comparing to Lowe's method[1], the proposed algorithm can deal with feature-less areas, providing therefore an environment-independent method for the image alignment task.

The following subsections describe the overall flow of processing. First, we briefly introduce the camera motion parametrization. Second, we introduce the global-to-local pairwise motion estimation, followed by the multi-view fine alignment description.

3.1 Camera Motion Parametization

Assuming that the camera undergoes purely rotations around it's optical center the camera motion can be parameterized by a 3×3 rotation matrix $\mathbf{R}$ and the camera calibration matrix $\mathbf{K}$. Under the pinhole camera model, a point in space $\mathbf{p} = (p_x, p_y, p_z)^T$ gets mapped to a 2D point $\mathbf{u} = (u_x, u_y)^T$ through the central projection process, which can be written using the homogenous coordinates $(u_x, u_y, 1)^T$ as following:

$$\begin{pmatrix} u_x \\ u_y \\ 1 \end{pmatrix} \cong \mathbf{KR} \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} \quad (1)$$

[1]In practice we may notice visible seams due to images' misalignment. One of the main reason is that the motorization of the capturing device yields some vibration noise which is further amplified by the tripod platform. Moreover, unmodeled distortions or failure to rotate the camera around the optical center, may result small amounts of parallax.

where, $\mathbf{K} = \begin{bmatrix} f & 0 & x_0 \\ 0 & f & y_0 \\ 0 & 0 & 1 \end{bmatrix}$ contains the intrinsic parameters, i.e. the focal f and the principal point offset (x_0, y_0). Note that pixels are supposed to be squares. Inverting equation 1 yields a method to convert pixel position to 3D-ray. Therefore, using pixels from an image (I_2) we can obtain pixel coordinates in another image (I_1) by projecting back into the I_1's space using equation 1. This principle can be summarized by the warping equation, which is expressed as:

$$\hat{\mathbf{u}}_1 \cong \mathbf{K}_1 \mathbf{R}_1 \mathbf{R}_2^T \mathbf{K}_2^{-1} \mathbf{u}_2 \quad (2)$$

Assume that all the intrinsic parameters are known and the same for all n images, i.e. $\mathbf{K}_i = \mathbf{K}, i = 1, .., n$. We choose unit quaternions for rotations parametrization, which are compact and elegant for numerical optimizations [6]. An unit quaternion is a normalized four-dimensional vector $\dot{\mathbf{q}} = (q_0, q_x, q_y, q_z)$. A rotation of angle ψ around an axis $\mathbf{n}$ can be represented by the unit quaternion $\dot{\mathbf{q}} = (\cos\frac{\psi}{2}, \sin\frac{\psi}{2}\,\hat{\mathbf{n}})$ where $\hat{\mathbf{n}}$ is the unit vector $\hat{\mathbf{n}} = \frac{\mathbf{n}}{\|\mathbf{n}\|}$. The orthogonal matrix $\mathbf{R}(\dot{\mathbf{q}})$ corresponding to a rotation given by the unit quaternion $\dot{\mathbf{q}}$ is expressed by:

$$\mathbf{R}[\dot{\mathbf{q}}] = \begin{pmatrix} q_0^2 + q_x^2 - q_y^2 - q_z^2 & 2(q_x q_y - q_0 q_z) & 2(q_0 q_y + q_x q_z) \\ 2(q_0 q_z + q_x q_y) & q_0^2 - q_x^2 + q_y^2 - q_z^2 & 2(q_y q_z - q_0 q_x) \\ 2(q_x q_z - q_0 q_y) & 2(q_0 q_x + q_y q_z) & q_0^2 - q_x^2 - q_y^2 + q_z^2 \end{pmatrix} \quad (3)$$

In order to handle deviation from the pure parallax-free motion or ideal pinhole camera model we improve the camera motion model by adding local motion estimation provided by a patch-based local matching procedure.

3.2 Global-to-local Pairwise Motion Estimation

The proposed framework starts with the global rotation motion estimation followed by the parallax compensation which is performed via a patch-based local motion estimation.

Rigid Rotation Computation. The global rotation estimation follows four steps: (i) pyramid construction, (ii) patch extraction, (iii) motion estimation and (iv) coarse-to-fine refinement. At every level of the pyramid l the goal is to find the 3D rotation $\mathbf{R}^l$. Since the same type of operation is performed at each level l, let us drop the superscript l through the following description.

Let $\mathbf{R}(\dot{\mathbf{q}}_\theta, \dot{\mathbf{q}}_\varphi, \dot{\mathbf{q}}_\psi)^{init}$ be the initial guess provided by the pan-tilt head, where θ, ϕ, ψ denote the pitch, roll and yaw angles respectively expressed in the camera coordinate system. The optimal rotation is computed by varying the rotation parameters θ, ϕ, ψ within an homogeneous *pyramidal searching space*, $\mathcal{P}_{\mathcal{SS}}$, which is recursively updated at each pyramidal level. For a given rotation $\mathbf{R}_{(\theta,\varphi,\psi)}, (\theta, \varphi, \psi) \in \mathcal{P}_{\mathcal{SS}}$ we can map pixels $\mathbf{u}_2^j$ from I_2 in the I_1's space using the warping equation expressed in equation 2. The optimal rotation is obtained by maximizing the similarity in brightness between $I_1(\mathbf{u})$ and $I_2(\mathbf{u}; \mathbf{R})$ in the overlapping region. The Zero Normalized Cross Correlation $(\mathcal{Z})$ is used as similarity measure[2] which provides robustness to illumination changes. The global similarity measure is given by the mean of all the similarity scores computed for all the patches belonging to the overlapping region.

$$\mathbf{E}[\mathbf{R}_{(\theta,\varphi,\psi)}] = \frac{1}{N_w} \sum_{j=0}^{N_w - 1} \Phi_j \mathcal{Z}(I_1(\mathbf{u}^j), I_2(\hat{\mathbf{u}}^j_{\mathbf{R}_{(\theta,\varphi,\psi)}})) \quad (4)$$

[2]computed on an integration window of size $(2w+1)\times(2w+1)$

Φ_j defines a characteristic function which penalizes "lost"[3] and "zero"[4] pixels and N_w denotes the number of valid matches belonging to the overlapping area. The optimal rotation $\hat{\mathbf{R}}_{(\theta,\varphi,\psi)}$ is obtained by maximizing the similarity score $\mathcal{Z}$ over the entire searching area $\mathcal{P}_{SS}$.

$$\hat{\mathbf{R}}_{(\theta,\varphi,\psi)} = \arg \max_{(\theta,\varphi,\psi)\in\mathcal{P}_{\mathcal{SS}}} \mathbf{E}[\mathbf{R}_{(\theta,\varphi,\psi)}] \quad (5)$$

Local Motion Estimation. We use the rotationally aligned images to perform the local patch matching. Let $\mathbf{P}_1 = \{\mathcal{P}(\mathbf{u}_1^k)|\mathbf{u}_1^k \in I_1, k = 1, ..., N_1\}$ and $\mathbf{P}_2 = \{\mathcal{P}(\mathbf{u}_2^k)|\mathbf{u}_2^k \in I_2, k = 1, ..., N_2\}$ be the patches extracted in image I_1 and I_2 respectively, which are defined by a neighborhood $\mathcal{W}$ centered around $\mathbf{u}_1^k$ and $\mathbf{u}_2^k$ respectively. For each patch $\mathcal{P}(\mathbf{u}_1^k) \in \mathbf{P}_1$ we search for its optimal match in I_2 by exploring a windowed area $\mathbf{W}^{\mathbf{SA}}(\mathbf{u}_2^k; \hat{\mathbf{R}})$ centered around $(\mathbf{u}_2^k; \hat{\mathbf{R}})$, where $\mathbf{SA}$ denotes the searching area ray. Let $\mathbf{P}_2^{k,\mathbf{SA}} = \{\mathcal{P}(\mathbf{u}_2^m)|\mathbf{u}_2^m \in \mathbf{W}^{\mathbf{SA}}(\mathbf{u}_2^k; \hat{\mathbf{R}}) \subset I_2, m = 1, .., M\}$ be the M patches found by exploring the searching area with 1-pixel steps. The best match is obtained by computing the similarity scores for each patch $\mathcal{Z}(I_1(\mathbf{u}^k), I_2(\mathbf{u}^m))$ and maximizing the score via bicubic interpolation in order to provide the best match with subpixel accuracy and real time performances. This yields a list of AF matches $(\mathbf{u}_1^k; \hat{\mathbf{u}}_2^k), k = 1, .., N$ and the possibility to compute a local translational motion for each match: $\mathbf{t}^k = \|(\mathbf{u}_2^k; \hat{\mathbf{R}}) - \hat{\mathbf{u}}_2^k\|$.

Figures 2 and 3 illustrate the results obtained by running the global-to-local image motion estimation procedure on an image pair gathered in the Tautavel prehistoric cave, France. The capturing device was set to acquire high resolution images of size 3008×2000 with an overlap of $\simeq 33\%$. In order to evaluate our technique with respect to a feature-based method, we show the results obtained on an image pair for which the SIFT detection and matching failed. The rotation computation starts at the lowest resolution level, $L_{max} = 5$ where a fast searching is performed by exploring a searching space $P_{SS}^{L_{max}} = 5°$ with 1-pixel steps in order to localize the global maximum (Fig. 2c). The coarse estimation is refined at higher resolution levels $l = L_{max} - 1, .., 0$ by taking a P_{SS} of 4 pixels explored with 1-pixel steps. Since deviations from parallax-pure motion are negligible we speed up the process by computing the local motion directly at the highest resolution level, $l = 0$ (Fig. 3).

After translation compensation, the camera motion consists in purely rotations. Therefore, the optimal rotation minimizes the angle between the corresponding 3D rays of each match pair. Table 1 illustrates the residual mean square error ($\bar{\mathbf{r}}$) and the standard deviation ($\sigma_{\mathbf{r}}$) of the pairwise camera motion estimation $[\hat{\mathbf{R}}, \mathbf{t}^k]$, computed with two different criterions: the projection error in the 2D space (equation 6) and the angle between the corresponding 3D-rays given by their cross product (equation 7). The second row of table 1 verifies the rotation estimation correctness showing that the angular distance $RMS_{\times 3D}$ between the non-aligned images (first column) corresponds to the optimal rotation estimate, $\hat{\mathbf{R}}$.

[3]the pixel falls outside of the rectangular support of I_2

[4]missing data either in $I_1(\hat{\mathbf{u}}_{\mathbf{R}}^j)$ or $I_2(\hat{\mathbf{u}}_{\mathbf{R}}^j)$, which may occur when mapping pixels $\hat{\mathbf{u}}_{\mathbf{R}}^j$ in the I_2's space

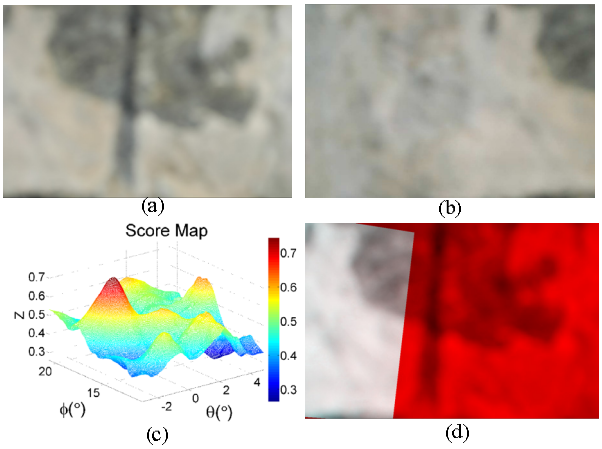

Figure 2: Rigid Rotation Estimation.(a)origin I_1, (b)image to align I_2, (c)global maximum localization at level $L_{max} = 5$, (d)rotationally aligned images at level $l = 0$: I_1-red channel, the warped image $I_2(\mathbf{u}; \hat{\mathbf{R}})$-green channel, $\hat{\mathbf{R}}(\theta, \varphi, \psi) = (17.005°, 0.083°, 0.006°)$.

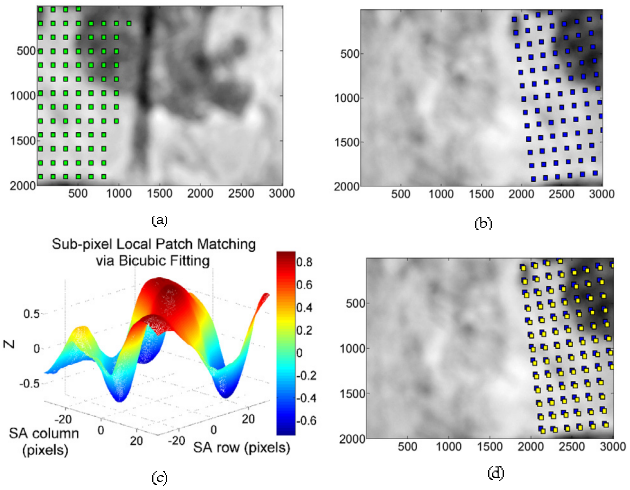

Figure 3: Anonymous Features Matching Procedure. $\mathcal{W} = 15$ pixels, 85 AF matches. (a)$\mathcal{P}(\mathbf{u}_1^k)$, (b)$\mathcal{P}(\mathbf{u}_2^k)$ extraction in I_2 using the rotation initialization, (c)Bicubic fitting for an arbitrary patch: $\mathbf{SA} = 32$ pixels, matching accuracy: 0.005 pixels, (d)AF-based optical flow: $\mathcal{P}(\mathbf{u}_2^k)$ blue, $\mathcal{P}(\hat{\mathbf{u}}_2^k)$ yellow, $\bar{\mathbf{t}} = [1.6141, 1.0621]$ pixels.

$$\bar{\mathbf{r}}_{2D} = \frac{1}{N} \sum_{k=1}^{k=N} \|\mathbf{u}_i^k - \mathbf{K}\hat{\mathbf{R}}_{ij}^T \mathbf{K}^{-1}(\hat{\mathbf{u}}_j^k - \mathbf{t}^k)\| \quad (6)$$

$$\bar{\mathbf{r}}_{\times 3D} = \frac{1}{N} \sum_{k=1}^{k=N} \|\mathbf{p}_i^k \times \hat{\mathbf{R}}_{ij}^T \mathbf{K}^{-1}(\hat{\mathbf{u}}_j^k - \mathbf{t}^k)\| \quad (7)$$

3.3 Multi-view Fine Alignment via Bundle Adjustment

Given the pairwise motion estimates $\hat{\mathbf{R}}_{ij}$ and the associated set of AF matches $\mathbf{P}(i,j) = \{(\mathbf{u}_i^k \in I_i; \hat{\mathbf{u}}_j^k \in I_j)|i \neq j, j > i\}$, we refine the pose parameters jointly within a bundle adjustment process [10]. This step is a critical need, since the simple concatenation of pairwise poses will disregard multiple constraints resulting in mis-registration and gap. As a first approach, we used the bundle adjustment framework described in [1], in which the objective function is a robust sum squared projection error. Given a AF correspondence $\mathbf{u}_i^k \longleftrightarrow \hat{\mathbf{u}}_j^k$ the error function is obtained by summing the robust residual errors over all images:

$$e = \sum_{i=1}^{n} \sum_{j\in I(i)} \sum_{k\in\mathbf{P}(i,j)} h(\mathbf{u}_i^k - \mathbf{K}\hat{\mathbf{R}}_{ij}^T \mathbf{K}^{-1}\hat{\mathbf{u}}_j^k) \quad (8)$$

Table 1: Residual Error Measures. $\hat{\mathbf{R}}(\theta,\varphi,\psi) = (17.005°, 0.083°, 0.006°), \bar{\mathbf{t}} = [1.6141, 1.0621]$ pixels

$\bar{\mathbf{r}} \pm \sigma_{\mathbf{r}}$	no model	$\bar{\mathbf{t}}$ compensation	$[\hat{\mathbf{R}}, \bar{\mathbf{t}}]$ model
RMS_{2D}(pixels)	1989.68 ± 62.83	1988.05 ± 62.74	0.08 ± 0.01
$RMS_{\times 3D}(°)$	16.99 ± 0.51	16.98 ± 0.5	$(7 \pm 1) \times 10^{-4}$

where n is the number of images, $I(i)$ is the set of images adjacent to image I_i and $h(\mathbf{x})$ denotes the Huber robust error function [7] which is used for outliers' rejection. This yields a non-linear least square problem which is solved using the Levenberg-Marquardt algorithm. A detailed description of this approach may be found in [1].

4 Results and Performance Evaluation

Figure 4 illustrates two examples of hemispherical mosaics obtained by running the proposed Gigapixel mosaicing procedure on several hundreds of images acquired in the Tautavel prehistoric cave, France. As shown in figures 4 (a) and (b), the bundle adjustment step minimizes a criterion measured in the 2D space, leading to mis-registration errors, which yields correct results if a sufficient number of AF matches are given. Our first concern is to improve the multi-view fine alignment process by simultaneously computing the optimal quaternions minimizing a criterion computed in the 3D space in order to reduce the residual error when using a minimum number of AF correspondences. We use the spherical projection within the rendering pipeline introduced by [1]. The mosaic's high photorealistic level is emphasized by a high-performance viewer which allows for mosaic visualization using 4-level of detail (LOD).

5 Conclusions and Future Work

This paper presents two main contributions: first, we introduce the *anonymous features* which are an environment-independent features and can be employed for image matching, tracking or localization purposes. Second, a *global-to-local* pairwise image alignment is proposed which combines the state of the art methods complementarity (direct vs. feature-based). The two ingredients are combined to propose an automatic Gigapixel mosaicing system for generating *in situ* photorealistic mosaics of previously unknown, complex and unstructured underground environments, without requiring human operator intervention. The proposed technique can deal with several key issues of the Gigapixel mosaicing problem in unstructured and large-scale environments, such as: handling the absence of reliably detectable and trackable features, robustness to noisy initial guess and to deviations from pure parallax-free motion, high resolution image alignment with real-time performances, and robustness to illumination changes and blur. We demonstrated the reliability of our method by automatically generating photorealistic mosaics of a challenging underground prehistoric cave. In order to decrease the number of matches required by the bundle adjustment step, improving the multi-view fine alignment is an ongoing work. In the near future, the produced Gigapixel mosaics will be made available on the world wide web[5] allowing for virtual visits of the Tautavel prehistoric cave, (France).

[5] www.iCaves.fr

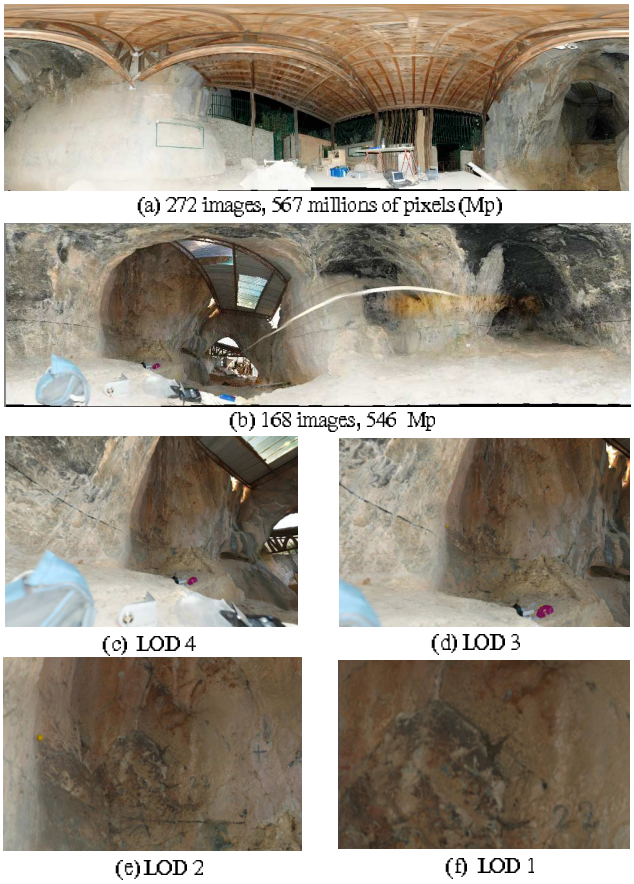

Figure 4: Mosaicing tests on data sets acquired in Tautavel prehistoric cave, France. (a) - cave's entrance (RMS_{2D}: 1.93 pixels, CPU: 8h 12 min), (b) - cave's center (RMS_{2D}: 1.76 pixels, CPU: 5h 33 min), (c), (d), (e), (f) - illustrate several LODs corresponding to the left part of mosaic (b).

Aknowledgements

The first author would like to thank to Alexandre Devaux for his help in using the Gigapixel mosaic viewer.

References

[1] M. Brown and D. G. Lowe. Automatic Panoramic Image Stitching using Invariant Features.*Int. J. Comput. Vision*,vol. 74, no. 1, pp. 59-73, 2007.

[2] S. E. Chen. QuickTime VR®: An image-based approach to virtual environment navigation. In *ACM SIGGRAPH '95*, pp. 29-38, 1995.

[3] N. Chiba, H. Kano, M. Higashihara, M. Yasuda, M. Osumi. Feature-based Image Mosaicing. In *IAPR Workshop on Machine Vision Applications*, pp. 5-10, 1998.

[4] S. Coorg and S. Teller. Spherical mosaics with quaternions and dense correlation. In *International Journal of Computer Vision*, vol. 37, pp. 259-273,2000.

[5] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, vol. 24, no. 6, pp. 381-395, 1981.

[6] B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, vol. 4, no. 4, pp. 629-642, 1987.

[7] P. J. Huber.*Robust Statistics*.Wiley,1981.

[8] Realviz.http://www.realviz.com

[9] R. Szeliski. Image alignment and stitching: A tutorial.*Foundations and Trends in Computer Graphics and Computer Vision*, Vol. 2, No. 1, December, 2006.

[10] W. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle Adjustment: A modern synthesis. In *Vision Algorithms: Theory and Practice*, no 1883 in LNCS, pp. 298-373.Springer-Verlag, Corfu, Greece, September 1999.