ClairVoyance: A Fast And Robust Precision Mosaicing System for Gigapixel Images

Frank Nielsen and Noriyuki Yamashita

Sony Computer Science Laboratories Incorporated
Fundamental Research Laboratory
Tokyo, Japan

Sony Corporation
Semiconductor Business Unit
System LSI Business Group
Tokyo, Japan

E-mail: Frank.Nielsen@acm.org and Noriyuki.Yamashita@jp.sony.com
http://www.csl.sony.co.jp/person/nielsen/

Abstract— **In this paper, we present ClairVoyance: a fully automatic photomosaicing system for building ultra high-resolution composited images from image sequences captured by tailored in-house motorized active pan-tilt digital camera units. Our stand-alone mobile systems built over the past five years are computationally fast, robust to various datasets and deliver unprecedented consumer-level image quality. We describe our simple yet novel lens calibration and radiometric correction procedures based on a fast block matching algorithm. All of our core image stitching components are based on the 2D Fourier phase correlation principle, and are thus easily amenable to hardware LSI implementation. We validate our approach by presenting sharp photomosaics obtained from a few hundreds up to a few thousands data sets of images.**

I. Introduction

Although consumer digital cameras are nowadays equiped with tele lenses that provide large optical zoom (×10 and more[1]), we still need to search and align *manually* the appropriate region of interest (ROI) **before** taking the picture. How nice would it be to take at first the full resolution picture, and then **later on** retrieve manually or *automatically* the possibly many regions of interests! Such a high-resolution picture, commonly called a photomosaic, better immortilizes a "moment." Photomosaicing has numerous applications: high resolution real-world capturing (*e.g.*, bring back in matter of seconds wonderdul sceneries one saw during his/her leisure trips), security, digital archiving, etc. Our system called ClairVoyance[2] can acquire large field of view photos with high resolution in a matter of seconds or minutes.

Section 2 gives a concise overview of prior work. Section 3 describes our hardware acquisition system and novel stitching algorithm. Section 4 reports on our implementation and presents photomosaic results. Finally, Section 5 concludes our paper.

II. Previous Work

Stitching or *mosaicing* pictures is nowadays a key ingredient of image processing. The first computer 2D stitching experiment was reported back in the mid 1970s. Since then, mosaicing theory and techniques have steadily been improved. Just to name a few corner milestones, let us cite the Fourier phase correlation mosaicing [1] in 1975 and the 360-degree cylindrical mosaicing [2] in 1995 that yielded Quicktime VR®. Since then, we have attested at a plethory of techniques for full spherical mosaicing [3] and full spherical video mosaicing [4] (4π steradians field of view). Mosaicing pictures allows either to increase the field of view (fov) to deliver panoramic imageries, or to increase significantly the image resolution (measured in dots per inch (dpi), useful for printing high quality posters). Besides, mosaicing is also a proven technology for signal-to-noise reduction (better SNRs through

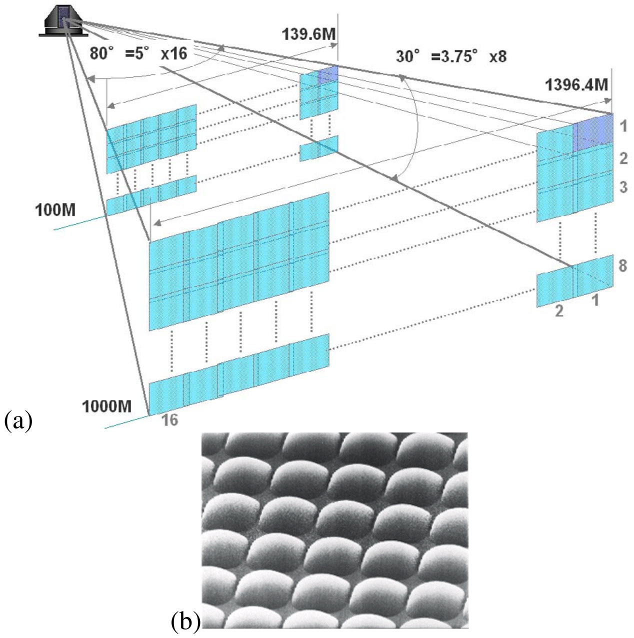


Fig. 1. Acquiring a large field of view image with high resolution by stitching many image tiles. (a) Example of a photomosaic consisting of $16 \times 8 = 128$ tiles covering a $80^o \times 30^o$ field of view. Each tile covering a $5^o \times 3.75^o$ field of view is captured using a consumer image sensor equiped with a tele lens. At 100 meters, the covered area is about 140 meters wide. That is, for 10.5 million pixel sensor we get 4.45 pixels/cm. (b) Each tile pixel is captured itself using an on-chip micro-lens (Super HAD CCD microscopic picture, courtesy of © Sony Corp.).

[1] This allows one to discover details at first not directly visible by Human eyes.

[2] Dictionary excerpt: Clairvoyance: n. A power, attributed to some persons while in a mesmeric state, of discerning objects not perceptible by the senses in their normal condition.

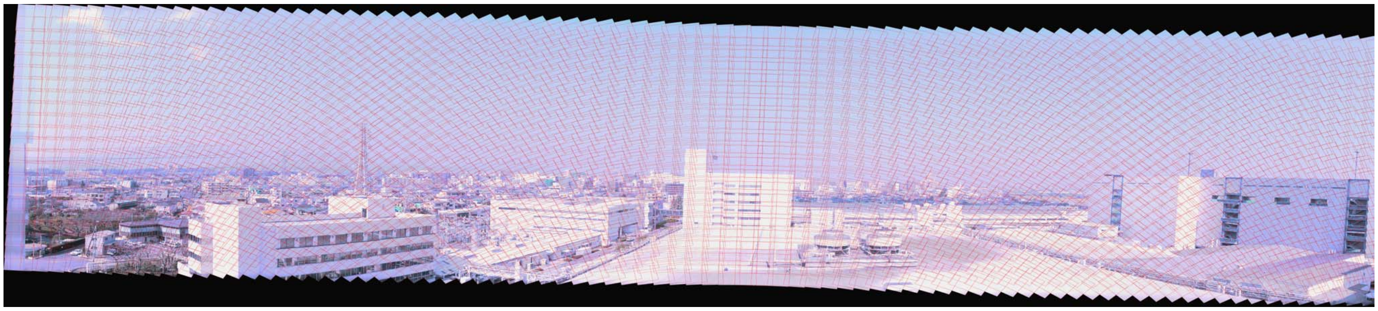


Fig. 2. Initial tile positions of our first capturing prototype that used a video camera with a motor-controlled mirror to acquire image tiles at 30 fps using a vertical scanline saccade (1800 tiles red bordered, captured in 2001). Tile positions are first regularly initialized before performing the block-matching process.

superresolution), high-dynamic range and multi-focus sharp images. With the advent of fast computing chips and cheap next generation disc storage devices (50 GB), it becomes possible to capture and interactively view and enjoy ultra high resolution images.

The Millennium map[3] offers the whole world imagery at one meter resolution (a gigantic 10,000 gigapixel image —10 terapixels— obtained after registering over 160,000 satellite images). Yet another project, named `GigaPxl`[4], uses expensive custom-built cameras to capture high-resolution images at once, without stitching. In this paper, we present a fast and robust consumer-level system, called CLAIRVOYANCE, for acquiring ultra high-resolution images using robust, precise and yet fast software mosaicing techniques. Most of the mosaicing methods differ in their geometric registration and radiometric color correction techniques (see [5] and references cited therein).

In the following section, we describe next our acquisition settings, the geometric calibration technique, the global registration procedure, and finally the radiometric correction scheme of CLAIRVOYANCE.

III. THE CLAIRVOYANCE MOSAICING SYSTEM

In this section, we first present our in-house capturing devices. Next, we recall the basics of the Fourier phase correlation principle, and its use for fast 2D block matching. Then, we describe the camera calibration, global registration and radiometric correction modules.

A. Capturing Devices

It is crucial to get a *mobile* acquisition device so that one can easily travel and acquire within a few minutes a ultra high resolution panorama image. We have built and tested several systems over the past fiver years that all rely on in-house motorized pan-tilt turntables, leveled by a tripod, on which are mounted either a digital still camera or a video camera with a tele lens (narrow field of view). The imaging sensor is mounted such as to minimize the parallax effect: that is, the rotation of optics is performed as close as possible around the focal point. We designed LSI circuits using programmable logic devices (PLDs) to control remotely the orientation of the camera head by a PC. The speed of acquisition, extrema of horizontal and vertical field of views, and camera parameters are customizable. Ideally, we would like to use the vertical scanline order to capture images. However this setting requires a large back rush when going from an image tile column to the next column. Thus, we opted for a vertical zigzag scan order. Figure 2 shows the 1800 raw image tiles placed roughly at their initial capturing positions before precise automatic registration.

B. Phase Correlation Principle

Stitching image tiles is necessary as their initial positions does not produce a correctly registered mosaic. That is, one can easily notice visible seams due to misalignements of tiles. One of the main reasons is that the motorization of the capturing head yields some vibration noise that are futher amplified by the tripod platform. In practice, we may observe misalignments up to 10% of the image width. Since we are going to stitch algother thousands of images, we first need to review the basic principle of stitching two images, and then motivate our choice of 2D block matching method. Recent stitching methods [5], [6] proceed by first calibrating the camera-lens system (mainly, focal length and radial distortions) and then find for each image tile a pure 3D rotation (roll, pitch and yaw attributes). For 360^o cylindrical or 4π spherical panoramas, constraints may further be applied so that the images match well everywhere including at their boundaries. Most of those modern methods are based on first extracting features from images and then matching them to calibrate the camera and retrieve individual image rotations. Those methods use the so-called RANSAC[5] procedure to eliminate outlier features so that only remaining matching inliers are further optimized numerically using the Levenberg-Marquadt optimization. These methods have two drawbacks for our concern: (1) first, they use randomization and therefore their running times vary significantly (two orders of magnitude) which is not good for circuitry LSI implementations, and (2) they rely on point feature extraction. This later point is a major drawback for stitching fully textured images (such as sky portions). Note that for ultra high resolution image,

[3] See `http://www.millennium-map.com/`

[4] See `http://www.gigapxl.org/`

[5] Random Sample Consensus, a method pioneered By Fischler and Bolles in 1981 [7], and nowaday commonly used in computer vision.

many images have potentially such fully textured areas. Thus, we need (1) a deterministic method that (2) does not require point features, and further (3) that is robust to small amount of noise. Such a method is the *phase correlation* technique based on Fourier analysis. Namely, the Fourier shift theorem: *Shifting the spatial function only changes the phase in the spectral domain*. That is, suppose we are given two functions f_1 and f_2, such that one is the translation of the other, say $f_2(x, y) = f_1(x + x_t, y + y_t)$. Then, we have the following spectral property: $F_2(u, v) = F_1(u, v) \exp(-2\pi i(ux_t + vy_t))$. Equivalently, we rewrite the former equation using the *cross-power spectrum* (CPS for short):

$$\underbrace{\frac{F_1(u,v)F_2^*(u,v)}{|F_1(u,v)F_2^*(u,v)|}}_{\text{Cross}-\text{power spectrum}} = \exp(2\pi i(ux_t + vy_t)), \qquad (1)$$

where F_2^* denotes the conjugate function of complex function F_2. Note that the spatial inverse of $\exp(2\pi i(ux_t + vy_t))$ is the Dirac impulse function $\delta(x_t, y_t)$ (see [6]). In practice, images are a bit noisy so that the cross-power spectrum of two images is not a perfect Dirac (see Figure 3). Thus it is better to first localize the peak of the cross-power spectrum and then retrieve the translation parameters (x_t, y_t) from it. Localizing the peak can be done using *subpixel accuracy* by fitting for example some parametric quadratic surface. Phase correlation is one of the oldest mosaicing method and an important technique for image matching with numerous applications such as tracking features. The phase correlation method speeds up significantly the search of the best cross-correlation. Indeed, finding the best translation vector (x_t, y_t) boils down to compute the Fourier transforms of images. For images consisting of n pixels, the CPS can be computed using the fast fourier transform algorithm (FFT) in $O(n \log n)$-time. Moreover, FFT routines have been finely optimized over the years, and even multi-core and GPU implementations are available. Although we presented the 2D translation case, the method can be extended to similitudes and affine transformations as well [8]. However, no such *shift theorem* is known for the perspective projection case which explains the tendency in computer vision for feature-based RANSAC matching methods for image mosaicing.

For mosaicing 2D planar images such as oil paintings or traditional Japanese kakejiku, we may apply the phase correlation as is (assuming orthographic projection of the tele lens). Otherwise, for any two pictures acquired from a same nodal point with the same tilt orientation but potentially different raw and pitch angles, we first need to convert the image into a cylindrical (θ, s) image (depends on the focal length f) as follows: $\theta = \arctan \frac{x}{f}$ and $s = \frac{y}{\sqrt{x^2+f^2}}$.

The focal length in pixel units can easily be recovered from the horizontal field of view (hfov) and image width as $f = \frac{\frac{\mathrm{width}}{2}}{\tan\frac{\mathrm{hfov}}{2}}$. Once the cylindrical coordinate conversion is done, we apply the phase correlation method to retrieve the 2D translation vector.

IV. Calibration by Block Matching

Our camera lens calibration procedure is computed robustly using the block matching primitive (peak retrieval of the CPS).

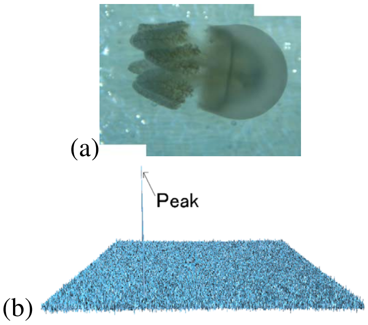


Fig. 3. Stitching images under translation using the cross-power spectrum. Image (a) shows the result of stitching two images. The 2D translation is obtained by detecting the peak of the cross-power spectrum (CPS), shown in (b). Here, the images perfectly match so that the CPS is readable. For noisy images, it is not that straightforward to detect the highest peak, we may rather consider a few candidate peaks to select from.

We first select a 2 × 2-tile region as depicted in Figure 4(a). The lens, horizontal and vertical fields of view (hfov and vfov) and tilt parameters are extracted from 2D block matchings. For each image tile border (say, $\mathrm{Picture}_i - \mathrm{Picture}_j$), we select three blocks in $\mathrm{Picture}_i$ and apply the fast phase correlation matching to find the corresponding block matching in $\mathrm{Picture}_j$. We get a vector displacement (x_{ij}, y_{ij}). By analyzing the behavior of the displacements of the three frontier blocks, we can quantify the importance of calibration parameters as follows:

- The effect of radial lens distortion is detected by the curve pattern of the matched block positions, as shown in Figure 4(c).
- The effect of the principal point displaced horizontally from the image center yields the pattern of Figure 4(d).
- The effect of image tilting is observed yet by another deformation pattern shown in Figure 4(e) and Figure 4(f).

Thus, we can calibrate the principal point (p_x, p_y), the radial lens distortions (we used two parameters κ_1 and κ_2) and the tilting angle of individual images from a simple, robust and fast block matching primitive. We repeat the calibration procedure on 2-tile regions until none of them have block displacements larger than a prescribed threshold.

V. Global Registration

Once the camera-lens calibration and tilting of each images is recovered from the local 2 × 2-tile optimization procedure, we considert the global registration. Global registration aims at finding the final positions of all image tiles at once so that the overall photomosaic is of best quality. Thus global registration extends the 2 × 2-tile region algorithm to the full image tile set. Given a reference image, we seek to find for all other images the 2D translations such that the overall registration error is minimized (Figure 4(b)). For spherical photomosaic, we fold/unfold locally images onto the sphere to perform the registration.

VI. Color Correction

Stitching without performing color correction yields photomosaic with noticeable color artefacts particularly visible in

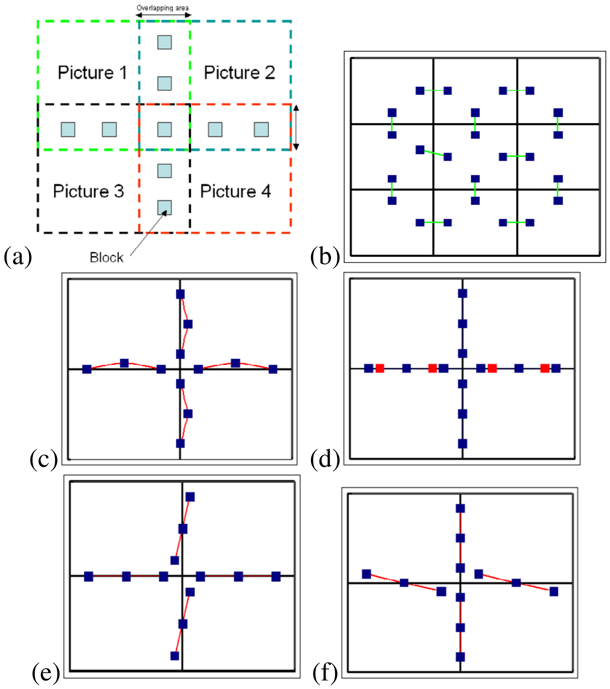


Fig. 4. Block matching: (a) overview of blocks in a 2 × 2-tile. Global registration

the sky area or other uniform textured parts. Color correction is a complex procedure that is computed for each color channel independently by the following pipeline: (1) Enhance image contrast, (2) Enhance image brightness, (3) Perform gamma correction, (4) Vignetting correction: Figure 5 explains the color correction procedure for a uniform sky part by plotting color graphs: Figure 5(a) shows the raw graph of three consecutive horizontal tile images. We observe the well-known law of squared cosine light falloff vignetting[6] effect [6]. Figure 5(b) displays the graph obtained after correction by a quadratic function. Observe that although each image color channel varies now linearly, the non-horizontal slope yields an overall global color gradiation. That is, although the sky color smoothly blends locally, we observe that the sky becomes progressively brighter when moving from one side extremity to the other. This artefact is caused by the principal point that does not coincide with the image center in practice. Thus in order to account properly for the lens vignetting phenomenon, we first need to recover the lens principal point, as described in Section IV. Once properly calibrated, the vignetting effect can be fully remove to yield a staircase graph, as shown in the plot of Figure 5(c). This last graph is easily adjusted by scaling uniformly and independently all images so that they match a given intensity level. All computations are done using 32-bit floating point arithmetic (high dynamic range image), and only at the last stage, shall we perform a simple tone mapping to get 8-bit/channel RGB photomosaic image.

(5) Global color adjustment: for each image tile, we consider its 8 direct neighbours and compute the respective difference of intensity levels. Adjusting locally the color by removing vignetting effects still yield to significant overall gradation problem (say, perceived on a photomosaic with 100 horizontal tiles). Thus, we define a gradation error function and scale independently all images so that the gradation function error is globally minimized. (6) Color blending: for each color channel, we blend each color pixel using a quadratic cross-fading function to produce a seamless composite picture.

[6] Informally speaking, pixels appear darker at the image corners.

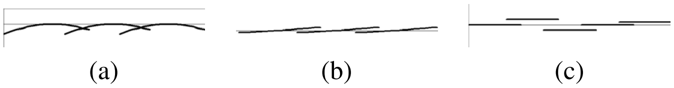


Fig. 5. Color correction graphs: (a) initial plot, (b) plot after basic vignetting correction, and (c) plot after vignetting correction using the principal point.

The process of (1) camera calibration, (2) global registration and (3) color correction can further be bootstrapped until the overall misregistration improvement fall within a prescribed threshold. In practice, we stop the iterations as soon as improvement gain falls below 1%. The last stage of CLAIRVOYANCE consists in trimming the stitched image to deliver a full rectangular image.

VII. CONCLUSION

The CLAIRVOYANCE system has been in daily used for gigapixel photomosaic acquisition since 2001. The current system's robustness has benefited from numerous data sets. In particular, we opted for a fully deterministic mosaicing approach that relies on a fast FFT block matching primitive to retrieve camera parameters, perform global alignment and radiometric correction. Figure 6 display a 400 million pixel Grand Canyon photomosaic (observe the quality of 102-tile radiometric correction). Figure 7 is an example of spherical photomosaicing obtained by over 3200 pictures that emphasizes on the clairvoyance functionality: by repeated zooming we may observe details that are at first not human visible. Figure 8 displays yet another outdoor planar photomosaic and an indoor spherical photomosaic. We are currently considering a pure LSI hardware implementation of CLAIRVOYANCE.

Acknowledgements

We would like to thank Dr. Hirofumi Sumi (Sony Corporation) for his precious help and valuable advices.

REFERENCES

[1] C. D. Kuglin and D. C. Hines, "The phase correlation image alignment method," in *Proc. IEEE 1975 Conference on Cybernetics and Society*, 1975, pp. 163–165.

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

[3] H.-Y. Shum and R. Szeliski, "Systems and experiment paper: Construction of panoramic image mosaics with global and local alignment," *International Journal of Computer Vision*, vol. 36, no. 2, pp. 101–130, 2000.

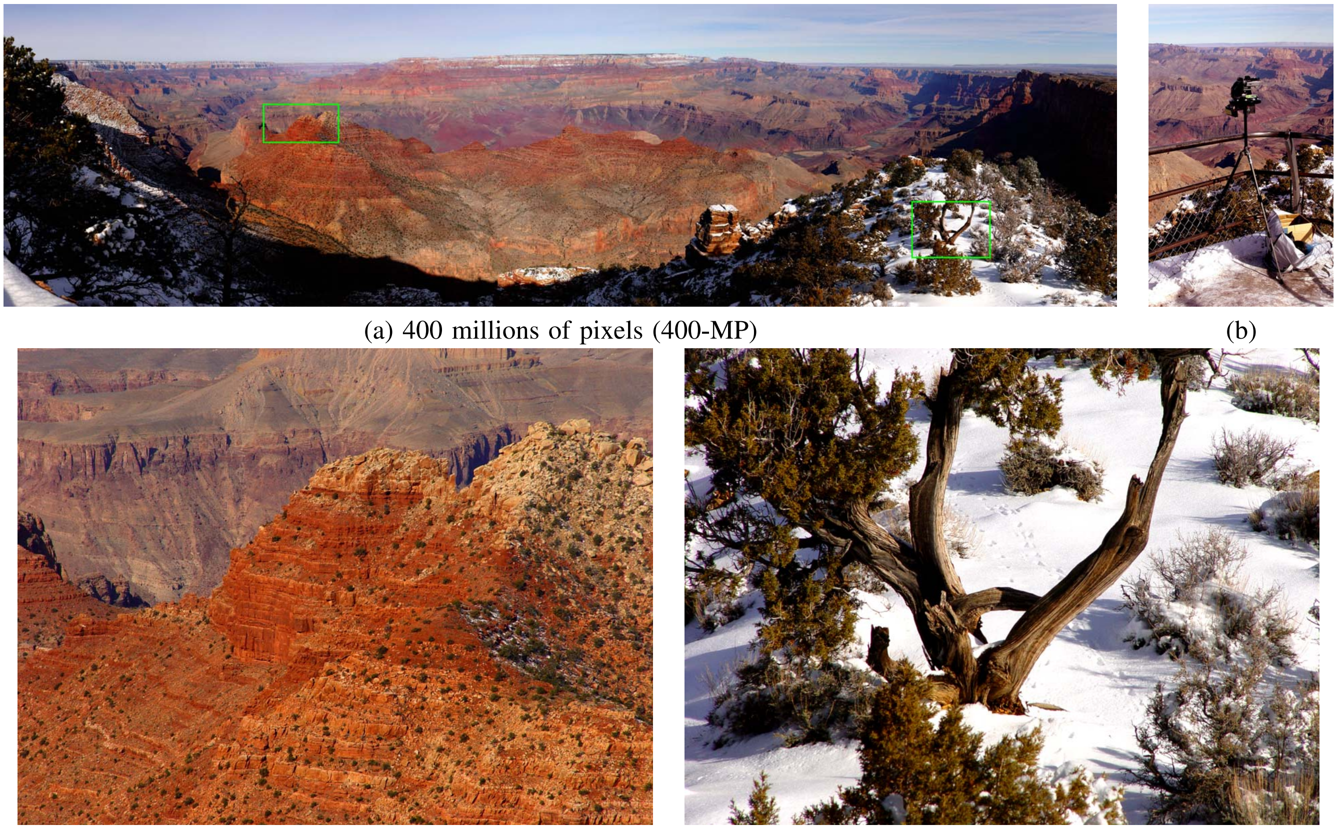
[4] F. Nielsen, "Surround video: a multihead camera approach," *The Visual Computer*, vol. 21, no. 1-2, pp. 92–103, 2005.

[5] R. Benosman and S. B. Kang, *Panoramic vision: sensors, theory, and applications*, Springer-Verlag New York, 2001.

[6] F. Nielsen, *Visual Computing: Geometry, Graphics, and Vision*, Charles River Media, ISBN 1584504277, 2005.

[7] 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.

[8] H. Shekarforoush, M. Berthod, and J. Zerubia, "Subpixel image registration by estimating the polyphase decomposition of cross power spectrum," in *IEEE CVPR*, 1996, p. 532



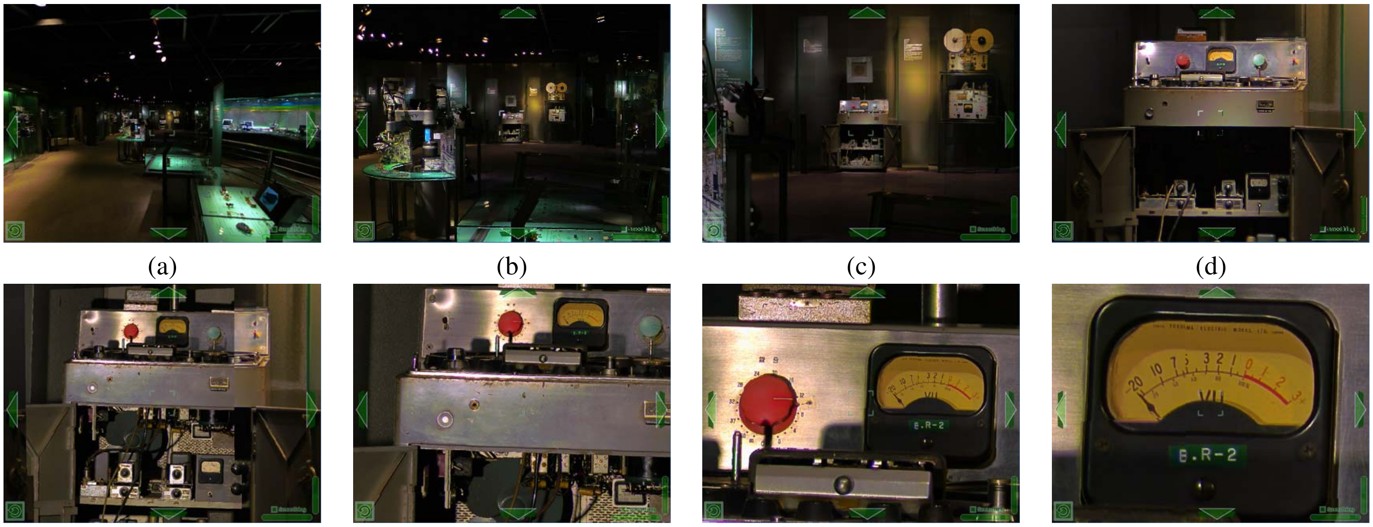
(a) 400 millions of pixels (400-MP)

(b)

(c) source picture (5-MP)

(d) source picture (5-MP)

Fig. 6. Grand Canyon photomosaic (UNESCO World Heritage #75, pictures acquired in 2003/03): (a) 37749 × 10253 ($\simeq$400 million pixels) high-quality $44.5^o \times 154^o$ cylindrical panorama (geometric registration error below 0.5 pixel and radiometrically corrected). (b) shows the mobile stand-alone acquisition system consisting of a Sony DSC-F717 digital camera controlled by an in-house motorized pan-tilt unit (capturing 2560 × 1920 images at 0.5 fps for an overall acquisition time of 4 minutes). Two of the $6 \times 17 = 102$ 5-MP source pictures are shown in (c) and (d), with corresponding windows in (a).



(a)

(b)

(c)

(d)

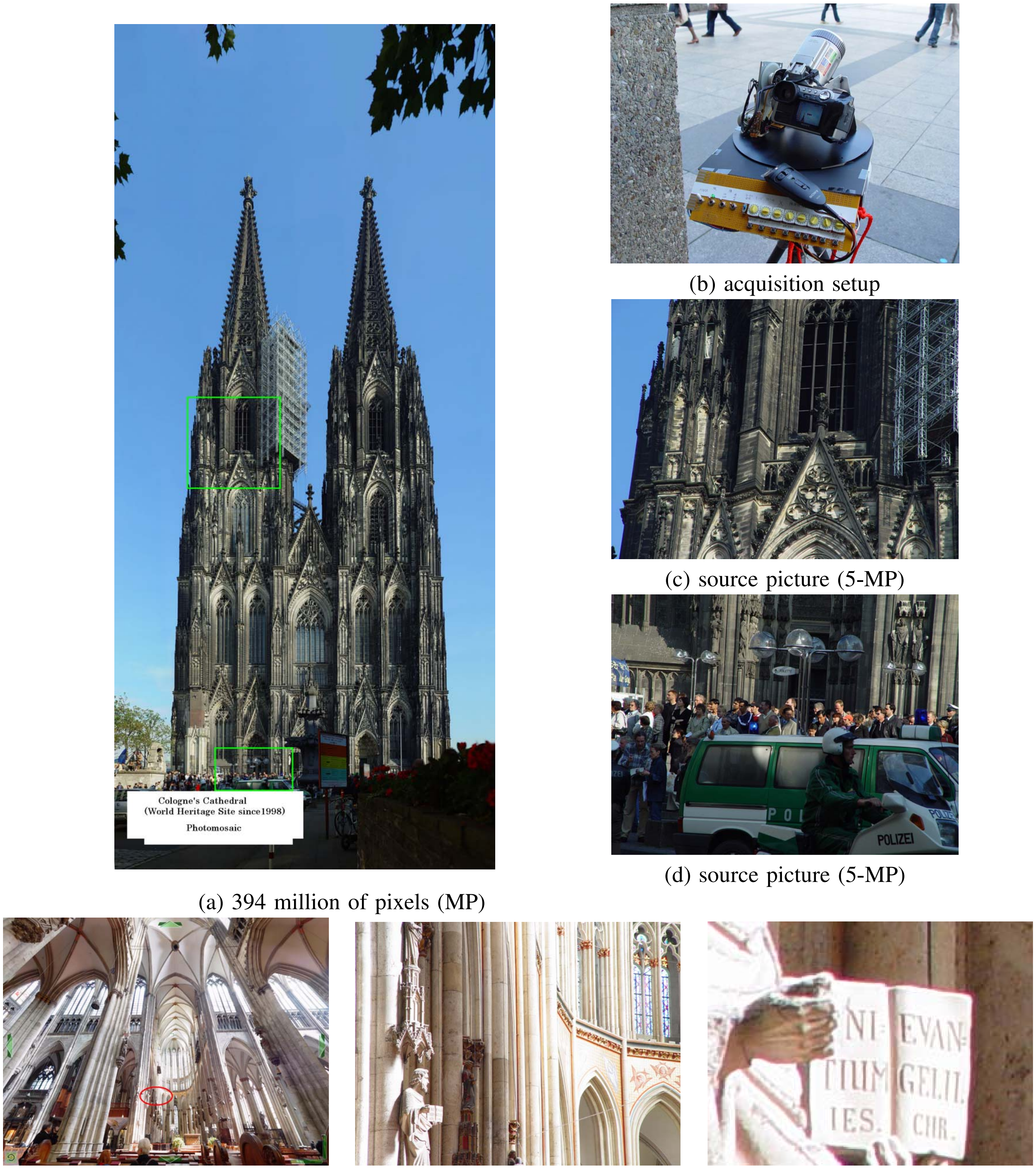
(e)

(f)

(g)

(h)

Fig. 7. Photomosaic of Sony Museum (Tokyo, Japan): A spherical gigapixel photomosaic acquired from over 3200 pictures by a Sony DFW-X700 digital video camera. (a) to (h) show snapshots of the interactive viewer session. (a) shows the tape recorder located 15 meters away from the nodal acquisition point (zoom level 0, horizontal field of view about 90 deg). (b) image after one 'step' zooming: the VU meter of the tape recorder becomes just distinguishable. (c) image at zooming level 3. (d) image at zooming level 4 (horizontal field of view 4 deg). (e) image at zooming level 5 (horizontal field of view: 1 deg) (f) and (g) zooming levels 6 and 7, respectively. (h) At zooming level 8, we can clearly read the VU meter of the tape recorder.



(a) 394 million of pixels (MP)

(b) acquisition setup

(c) source picture (5-MP)

(d) source picture (5-MP)

(e) viewer

(f) source picture

(g) close-up

Fig. 8. Cologne Cathedral exterior and interior photomosaics (UNESCO World Heritage #292rev, pictures acquired in 2002/09): (a) A $13704 \times 28800 \simeq 394$-MP photomosaic obtained from $13 \times 7 = 91$ pictures $(76\,\mathrm{deg} \times 53\,\mathrm{deg})$ revealing the fine details of the 150-meter height gothic cathedral (construction began in 1248 and finished in 1880). (b) shows the portable acquisition system that consists of a consumer digital camera (Sony DSC-F707) mounted on our in-house motorized pan-tilt table. (c) and (d) are two 5-MP source pictures with corresponding regions displayed in the photomosaic. (a) was printed on fifteen A3 sheets at 400 dpi resolution. (e) is a screenshot of our composited picture viewer showing the Cathedral's fully spherical interior $(3 \times 9 \times 11$ pictures acquired with tele lenses). (f) shows a source picture with its corresponding region in the viewer (e). (g) is a close-up emphasizing on the quality of source pictures (latin words on sculpted book).