Real-Time Spherical Videos from a Fast Rotating Camera

Frank Nielsen[1], Alexis André[1,2], and Shigeru Tajima[1]

[1] Sony Computer Science Laboratories, Inc.
3-14-13 Higashi Gotanda, 141-0022 Shinagawa, Tokyo Japan
frank.nielsen@acm.org
[2] Tokyo Institute of Technology
2-12-1 Oookayama, 152-8552 Meguro, Tokyo Japan

Abstract. We present a new framework for the acquisition of full spherical videos using a high-speed rotating linear or area sensor equipped with a fish-eye lens. The high capture rate of the sensors allows us to reach high frame rates on the resulting videos. We describe the image processing workflows from the raw data to the full spherical frames, and we show experimental results of non-parallax panoramic videos covering the complete field of view.

1 Introduction and Prior Work

Full spherical images have a wide range of applications from environment mapping [4] to virtual reality [3], but the field of full spherical videos is much more promising. In this paper, we focus on the creation and some applications of high frame rate full spherical videos acquired from a rotating video camera.

Since the beginning of art, Human civilizations have explored the possibilities of various fields of view when depicting daily or ritual scenes, for example with fresco paintings. We have to wait until the mid-18th century to find the first artistic full cylindrical panoramas, by the Irish painter Barker [1]. These first cylindrical panoramic paintings were popular and could be viewed in dedicated exhibition spaces, called cycloramas [1]. Although computers were already used to seamlessly *stitch* images in the mid-1970's [2], it is certainly the acclaimed work of Chen [3] in 1995 that leaded Apple to release the very first commercial hit product for authoring and viewing cylindrical panoramas in the digital arena: Quicktime VR®. This digital panoramic image processing workflow was further extended quickly to full field of view panoramas, at first by authoring a few dozens of pinhole images [4].

The laborious acquisition process was then improved and simplified by considering registering a few images (i.e., typically 2 or 3 images) obtained from a fish-eye lens camera attached to a tripod [5]. Nowadays, this method is still the prominent way of acquisition way for amateur photographers to get high-quality seamless non-parallax full spherical images.[2] Another popular approach

[1] http://www.riesenrundgemaelde.at/e/thema/kunst1.htm
[2] E.g., see Realviz's image stitcher dual shot mode, http://www.realviz.com/

A. Campilho and M. Kamel (Eds.): ICIAR 2008, LNCS 5112, pp. 326–335, 2008.

prone by professional photographers working in the Computer Graphics Industry (CGI) consists in using a slowly stepmotor-controlled rotating line sensor camera. These tailored systems allow one to get high-resolution High-Dynamic Range (HDR) panoramic images but are *inherently* slow because of the use of step motors. For example, the company Roundshot (`www.roundshot.ch`) proposes a 360-degree "livecam" system that updates the panoramic image every few minutes. Another company, SpheronVR (`http://www.spheron.com/`), built the SpheroCam HDR system for acquiring HDR spherical images using stepmotors. However, the system requires 30 seconds up to several minutes depending on the exposure time and type of lens used.

Thus acquiring full field of view high-resolution video-rate panoramic media is still nowadays a very challenging and open problem. One reason is that monolithic systems that consist of a single area sensor coupled with, say, a paraboloid mirror yield incomplete and irregularly sampled imageries (see catadioptric systems [1]). Another strategy consists in using an array of cameras *more or less* aligned at the same optical nodal point to capture the surrounding environment. High quality panoramic videos were successfully obtained for cylindrical fields of view by using pyramid-faceted mirrors to align virtual nodal points of respective block cameras. One of the very first system was proposed by Nalwa and is currently sold by the FullView[3] company. This multi-camera approach was extended to full spherical videos by aligning ten wide angle fields of view camera images acquired synchronously [6] in Sony's FourthVIEW system. Although the FourthVIEW system yielded the very first full spherical 30 fps panoramic videos in 2000, it suffered from inherent parallax problems. The parallax becomes quite noticeable if the captured dynamic objects of the scene are below a so-called parallax clearance threshold. That threshold is typically of the order of several meters and thus limits significantly its use. The second inherent difficulty in handling such a camera cluster approach is the radiometric corrections of individual images.

Thus, although the concept of spherical videos is now widely accepted as a commodity in the research community[4], its inherent image quality limitations (parallax, radiometric corrections) are still the bottleneck of making it a major video consumer medium, notwithstanding its high selling price. Another approach to get panoramic video content is to synthetically create it. For example, Agarwala et al. [7] described a graphcut algorithm to artificially *synthesize* such panoramic videos by stitching both in space and time domain image patches obtained from a slow rotating camera. They demonstrated their prototype for wide but not complete cylindrical fields of view only at very low frame rate.

In this paper, we investigate a novel approach based on a *high-speed rotating/ high frame rate* camera equipped with fish-eye lens to capture full spherical frameless "videos." The essential differences of our approach compared with low-speed rotating cameras controlled by step motors is that:

[3] `http://www.fullview.com`

[4] See the commercial `Ladybug2` package provided by PointGrey, `http://www.ptgrey.com/products/spherical.asp`

1. we do not know *a priori* precisely rotation angles for a given acquisition set, and
2. we cannot afford to reverse the rotation direction in order to come back at the origin position (rewind).

Stepmotor systems namely do that: they rotate 2π by small precision increments and then rotate back to the original position. This latter point requires us to solve for a new connectic in order to avoid cable twists and jams. We solved this problem by using *slip rings*, rotating electrical connectors on which we transmit video signals using the gigabit Ethernet protocol. Since the camera is rotating fast, say ideally at 1800 revolutions per minute (rpm — i.e., 30 fps) the spherical "images" are perceived from the retinal persistence of vision property. We studied such systems for the following two scenarii of line/area sensors:

– For line sensors, we do not need to align the nodal point with the rotating axis as every 360-degree round brings back the camera to its original configuration, and thus produces a *smooth spherical image* that is visually flawless (we merely juxtapose vertical strip lines). However, we stress out that this image does not coincide with the panoramic image obtained from a virtual full spherical lens camera (i.e., it does not have the property of a unique Center Of Projection, COP). In fact, we may even use this 2π round invariant property to acquire video-rate *stereoscopic panoramas* by shifting the nodal point off the rotation axis, as first suggested by Peleg [8].
– For area sensors, we can precisely calibrate the camera-lens system so as to align the nodal point with the rotation axis. Since we are demultiplexing time/space, and may interpret the single camera system at constant speed as a multi-head camera for which we have precisely aligned the nodal points. Therefore, high-speed rotations allow us to bypass the physical limitations of manipulating a camera cluster [6] but introduces other challenging image processing problems such as horizontal motion blur.

The paper is organized as follows. We briefly describe the hardware prototypes we built for acquiring panoramic videos with a complete field of view, reviewing for each category of device the image processing workflows, and point out the system limitations as well as novel challenges.

2 Line Sensor Platforms

2.1 Off-the-Shelves HDV Camcorders

The first prototype is designed to *conveniently* capture colorful environments with few changes such as natural phenomena (sunsets, for example). It consists of a consumer electronics video camera (Sony HDV1), delivering a *compressed* MPEG stream (MPEG2-TS, 1080i at 25 Mbps). The camcorder, equipped with a common fish-eye lens (Raynox) is fixed on top of a battery-powered rotating plate so as to get the maximum vertical resolution (i.e., we rotate the camcorder

Fig. 1. The first prototype uses an off-the-shelf camcorder equipped with a fish-eye lens mounted on a fast rotating platform to acquire real-time full spherical surround images. Surround images are delivered real-time over the network via a WiFi connection.

90 degrees for vertical strips of size 1440 pixels). This stand-alone mobile system can be brought easily to public places such as parks. Each 2π revolution requires approximately 30 seconds to two minutes to capture. The spherical image is created by juxtaposing the middle vertical line (i.e., camera horizontal line) so as to produce a smooth surround image band, where the horizontal axis is a mapping of the rotation angle. This naive stitching can be performed online and streamed on a network by connecting the camera-lens system to a small PC via IEEE1394 interface (Sony VAIO-UX connected to the camcorder by Ilink and streaming images via WiFi connection). Figure 1 displays the acquisition platform as well as the result of an indoor full spherical image. At acquisition time, the 1-pixel width bands are first stored successively in a raw binary format. We then use a fast image alignment technique to be described next to find the period width of each revolution. This then allows us to convert the frameless raw data into conventional 2D latitude-longitude environment map frames, and to compress these 2D surround images using legacy compression methods (e.g., JPEG2000 or MPEG4).

The obvious limitations of the system are its low speed acquisition that restricts it to very particular settings (i.e., day summary), and the image quality that suffers from MPEG2 compression.

2.2 High-Speed Rotating CCD Line Sensors

Keeping in mind the rationale behind the first prototype, our central motivation for the second prototype is to *increase* the frame rate and test whether we can physically reach the barrier of 1800 revolutions per minute, leading the path to 30 fps full spherical videos. To overcome the compressed nature of the images and the fact that we only used the central band in our first prototype, we selected a high-frame rate black & white CCD *line sensor* from manufacturer DALSA. The Nikor fish-eye lens we used gives a sharp imagery but had the drawback of being heavy and filling only one third of the line sensor (the vertical resolution is about 400 pixels.) The camera can acquire technically up to 67000 lines per second, but we used only 10% of this performance given the exposure conditions

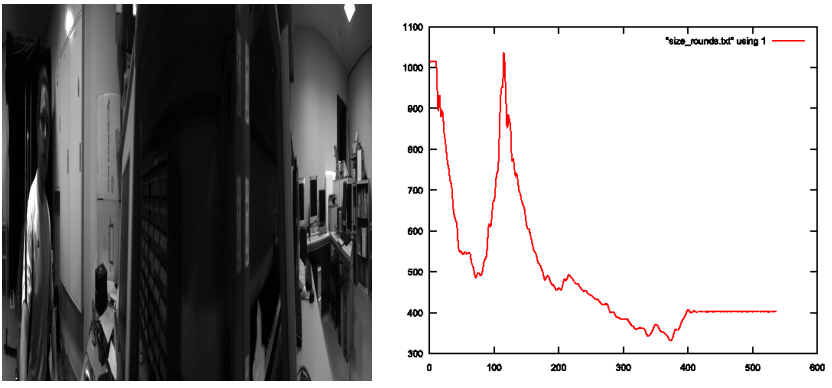

Fig. 2. Left: One frame extracted using the line sensor device ; the vertical resolution is fixed, while the horizontal resolution reflects the rotation speed. Right: Plot of the number of strips per round (angular speed) derived from strip alignments: the speed was manually controlled using a potentiometer as reflected in the graph fluctuations.

of our various shooting scenes. The main technical difficulty was to build a fast and stable rotating platform which speed is controlled manually by tuning a potentiometer. We used the DALSA GigE interface (Gigabit ethernet) and connected it to a slip ring. The camera is powered by lithium batteries located on the rotating platform. Although the slip ring is guaranteed for 400 rpm, we checked that it correctly transmitted the video signal up to 1000 rpm or even more. The software prototype uses the dedicated C++ camera SDK, and we used OpenGL® for previewing the frames in real-time. Overall, this second prototype was successful in proving that we could indeed target the 1800 rpm bound using slip rings for carrying video signals on Gigabit Ethernet.

2.3 Image Processing Workflow

We concisely explain the image processing tasks for stitching and authoring these frameless panoramas, and emphasize on the novel stitching tasks. For the *linear sensor* prototypes, the stitching process consists merely in writing one vertical band after another. The main challenge is then to assign an horizontal angle θ_i to each 1-pixel width vertical band b_i. We split the first "panoramic video band" into "panorama rounds" using the FFT-based Cross-Power Spectrum (CPS) technique [9] that runs fast, looking for the first repetition in the band. The CPS technique looks for a peak in the Fourier transform of the correlation of two images with significant overlap. The location of the peak corresponds to the best shift that transforms the first image into the second one. We use this technique by looking for the best shift of our panoramic band with itself. The best result now corresponds to the rotation period, if the rotation period is stable enough during the first revolutions.

We then assume that the first revolution is acquired at constant speed so that vertical strips of the first round are equally spaced in $[0, 2\pi)$. Then we register

all other vertical strips by using per-pixel sum-of-square-difference (SSD) local optimization, looking for the closest band in a neighborhood of the previously registered band in the revolution before. The most important band is the starting band, as it marks the starting point of a new revolutions. As the speed may change during one revolution, we actually register all bands. While the method is prone to error when parts are moving, we assume that a wide part of the scene is static. For our test scene, we were able to detect correctly the starting point of every revolution, with one person moving in the scene. Figure 2 reports on the alignment result for a data set with varying speed (manually controlled using the potentiometer).

2.4 Frame Coherency

One important issue in the described framework is the coherency between consecutive frames. The line-sensor based prototypes are in a sense robust to this problem, as the data used is not warped along the rotation direction, giving the same data revolution after revolution. However, since the speed may change between two revolutions, the resolution of the frames changes along. Moreover, if the speed present strong changes during one revolution, the resolution between different parts of the frame is different. One solution to this problem would be to register all bands with respect to the first revolution, and resize each complete revolution using this mapping as a non-linear weighting function, but this implies that the scene is static, which is not interesting.

On the contrary, if we suppose the rotation speed constant, the frames present the same resolution, and the resulting video does not show registration artifacts. In such cases, we achieve time coherency of the frames. In practice, the platform we built did present small variations in speed within one revolution, but on the whole, the speed of each revolution was constant.

3 High-Speed Rotating CMOS Area Sensor

3.1 Hardware Description

While the previous setup is able to capture high frame rate videos, two significant limitations are present: the image is only black-and-white, and the fisheye only covers one third of the whole scene. We therefore decided to build the last prototype for this project: full lens circle color image matching the resolution of the camera. Our third and current prototype uses a high frame rate CMOS color area sensor (Prosilica GC640C). It allows to capture roughly VGA size (659×493 pixels) Bayer-tile color picture at 200 frames per second. We may further increase the frame rate if we select smaller region of interests. For example, the frame rate hits 1000 fps for 100×100 pixel area. We replaced the Nikor lens by a light tailored C-mount fish-eye lens exploiting the full sensor dimensions. Figure 3 presents the hardware configuration.

Fig. 3. High-speed rotating motor with high-frame rate CMOS area color sensor. The right image is an example of a full spherical stitched picture band that shows several revolutions.

3.2 Stitching in Spacetime Domain

For the area sensor prototype, we consider the traditional pipeline [4,5,6] that proceeds by warping and stitching images into the latitude-longitude environment map (see Figure 4). We first "defish" the raw images [5]. For this purpose, we have to establish first a correct fisheye model represented by the relation between the distance to the optical center with the ray angle. We used a classic polynomial model, and we used a Levenberg-Marquadt optimization on a couple of test data, consisting in sets of feature points located in one global scene taken with the camera under different angles. Once the images are defished, we align them using global FFT CPS followed by a Levenberg-Marquadt SSD local minimization, giving the estimated angle between two consecutive frames. Warped and aligned images are then blended on-the-fly altogether using a simple polynomial blending rule, mixing the color components of two consecutive frames around the estimated seam. Poisson image blending [10], fusioning the gradient of the following frame at the junction between two consecutive frames, gives smoother results, in particular with respect to illumination changes, but the processing time required to solve the equation system makes it difficult so far to reach real-time blending.

3.3 Frame Coherency

The area sensor with the fisheye raises various issues when looking at the coherency of the frames. The main problem is the fisheye transformation, sampling the data with varying resolution. After registration, two consecutive frames are expected to contain the same information, but in practice, after "defish", some illumination artifacts often appear. While the Poisson-based image fusion is able to smoothen most of the artifacts out, the seams, while not obvious on a static frame, are visible on the resulting videos, as they are moving around the frame.

The problem is coming from a non perfect fish-eye model, and from saturated values, especially from light sources. The resulting haloes are very different according to their position on the captured frame, resulting in visible artifacts around them.

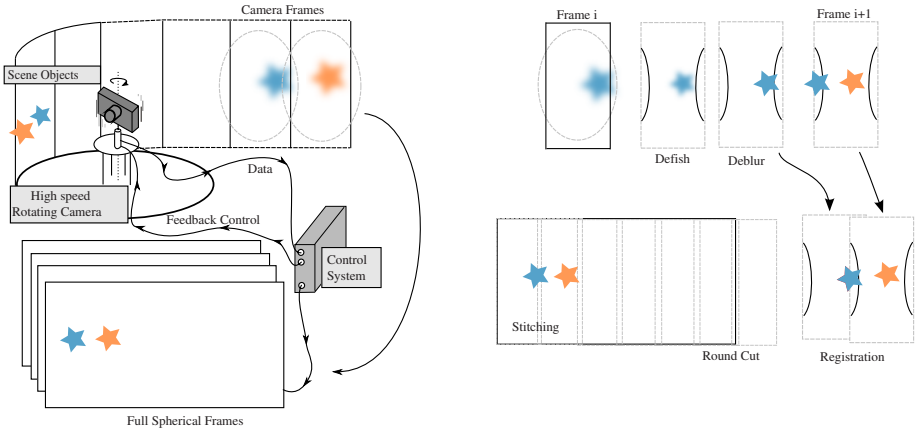

Fig. 4. Image processing workflow for the area sensor

3.4 Motion Blur

The main drawback of using area sensors in this project is the inherent presence of large horizontal motion blur. We use a small red LED box as depicted in

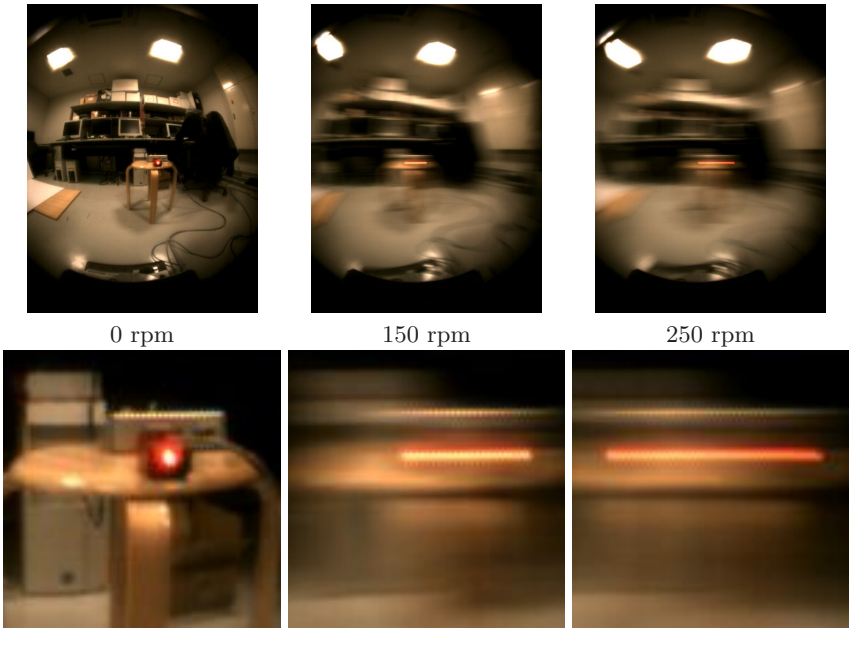

Fig. 5. Horizontal motion blur at different angular speeds (top). Close-ups of the red LED box emphasizing the horizontal point spread function.

Figure 5 to retrieve easily the *Point Spread Function* (PSF). While traditional deblurring methods (Lucy-Richardson or recent variational optimization algorithms [11]) are able to recover correctly the original image with the knowledge of the PSF, for extreme cases such as ours, we did not manage to recover a nice image. This blur is a real problem, as it also affects the performance of the registration process. The CPS technique we use is really sensitive to illumination changes, and the blur creates asymetric images (the overlapping part of two consecutive frames is not equally blurred due to the boundary conditions). We try to cope with this problem by first supposing the rotation speed constant, and performing double checks using a slower SSD when the CPS result is not close to the initial value.

4 Conclusion and On-Going Work

We presented in this paper a novel architecture for acquiring frameless panoramic videos by stitching in *spacetime* domain using high frame rate high-speed rotating cameras. Using slip-rings to transmit the data, we showed that it is physically feasible to reach full spherical panoramic videos at high frame rate using only one camera. Although our systems are capable of capturing full spherical panoramic videos, the main problems to cope with are the exposure time and/or the horizontal motion blur induced by the camera on-board imaging unit. If we consider the line sensor device, we succeeded in capturing blur-free cylindrical video-panoramas, but the current lens and sensor limit the captured scene to one third of our desired full sphere. The area sensor is however capable to capture the whole sphere, but is introducing motion blur, difficult to recover from.

One registration error is enough to ruin the whole video, so for real-world applications, a more robust method might be needed. We are currently investigating other sensors that allow shorter exposure times while keeping a reasonable image quality. The high rotation speed of the systems unavoidably yield small fluctuations. This gives us an opportunity for considering super-resolution using sub-angular pixel registration. Super resolution yet requires to consider another challenging task for our topic: motion detection and compensation.

The two sensors we used in this paper can be controlled remotely via the Ethernet link, allowing live changes in critical parameters such as exposure or region of interest, leading the way to the acquisition of HDR images, or content-dependent resolution.

Experiments and video samples are available online at: `http://www.sonycsl.co.jp/person/nielsen/spinorama/`

References

1. Benosman, R., Kang, S.B.: Panoramic vision: sensors, theory, and applications. Springer, Heidelberg (2001)
2. Milgram, D.L.: Computer methods for creating photomosaics. IEEE Trans. Comput. 24(11), 1113–1119 (1975)

3. Chen, S.E.: Quicktime VR: An image-based approach to virtual environment navigation. In: Proc. 22nd Computer graphics and interactive techniques (SIGGRAPH), pp. 29–38 (1995)
4. Szeliski, R., Shum, H.-Y.: Creating full view panoramic image mosaics and environment maps. In: Proc. 24th Computer graphics and interactive techniques (SIGGRAPH), pp. 251–258 (1997)
5. Xiong, Y., Turkowski, K.: Creating image-based VR using a self-calibrating fisheye lens. In: Proc. Computer Vision and Pattern Recognition (CVPR) (1997)
6. Nielsen, F.: Surround video: a multihead camera approach. The Visual Computer 21(1) (2005)
7. Agarwala, A., Zheng, K.C., Pal, C., Agrawala, M., Cohen, M., Curless, B., Salesin, D., Szeliski, R.: Panoramic video textures. In: ACM Trans. Graph (SIGGRAPH), pp. 821–827 (2005)
8. Peleg, S., Ben-Ezra, M., Pritch, Y.: Omnistereo: Panoramic stereo imaging. IEEE Trans. Pattern Anal. Mach. Intell (TPAMI) 23(3), 279–290 (2001)
9. Nielsen, F., Yamashita, N.: Clairvoyance: A fast and robust precision mosaicing system for gigapixel images. In: IEEE Industrial Electronics (IECON), pp. 3471–3476 (2006)
10. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Trans. Graph. 22(3), 313–318 (2003)
11. Fergus, R., Singh, B., Hertzmann, A., Roweis, S.T., Freeman, W.T.: Removing camera shake from a single photograph. ACM Trans. Graph. 25(3), 787–794 (2006)