High-Resolution Interactive Panoramas with MPEG-4

Peter Eisert, Yong Guo, Anke Riechers, Jürgen Rurainsky

Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute
Image Processing Department
Einsteinufer 37, D-10587 Berlin, Germany
Email: {eisert,guo,riechers,rurainsky}@hhi.fhg.de

Abstract

We present a system for the interactive navigation through high-resolution cylindrical panoramas. The system is based on MPEG-4 and describes the virtual world by the scene description language BIFS. This allows the easy integration of dynamic video objects, 3-D computer models, interactive scene elements, or spatial audio in order to create realistic environments. The scene data can be stored locally or streamed from a server dependent on the navigation information of the client. For the acquisition of panoramic views from real scenes, many preprocessing steps are necessary. Methods for the removal of objects or the adaptation of the dynamic range of the images are presented in this paper.

1 Introduction

Cylindrical panoramas for the creation of synthetic views from real scenes have a long tradition. Already in 1792, the painter Robert Barker built a panorama with a radius of 20 meters. Animated panoramas were presented around hundred years later in 1897 by Brimoin-Sanson. 10 synchronized projectors created the illusion of being present in foreign countries or distant places. Today, people are still attracted by gigantic panoramas like Asisi's 36 meters high Mount Everest panorama.

In image-based rendering [1], cylindrical panoramas have received particular interest in current applications due to their simple acquisition setup. Only a couple of pictures need to be captured on a tripod or freely by hand [2]. The images are stitched together forming one panoramic image as shown in Fig. 4. From the 360° scene information, new views can be rendered which enables the user to turn the viewing direction and interactively decide the point of interest. One well known example for such a system is QuicktimeVR [3].

Figure 1: Multiple panoramas from the Ethnological Museum in Berlin. Interactive scene elements allow the user to jump between the rooms. Dynamic objects are added to vitalize the scene.

In contrast to light fields [4] or concentric mosaics [5], the viewing position for panoramic rendering is restricted to a single point. Only rotation and zoom are permitted for navigation. This restriction can somewhat be relaxed by allowing to jump between different panoramas as shown in Fig. 1. However, for many applications this is sufficient and panoramic views can be found more and more often on web sites creating virtual tours for city exploration, tourism, sightseeing, and e-commerce.

In this paper, we present a system for streaming and rendering of high-resolution panoramic views that is based on MPEG-4. The use of MPEG-4 technology provides many new features compared to conventional 360° panoramas. Video objects, dynamic 3-D computer models [6, 7], or spatial audio as illustrated in Fig. 2 can be embedded in order to vitalize the scene. Pressing interactive buttons gives additional information about objects or modi-

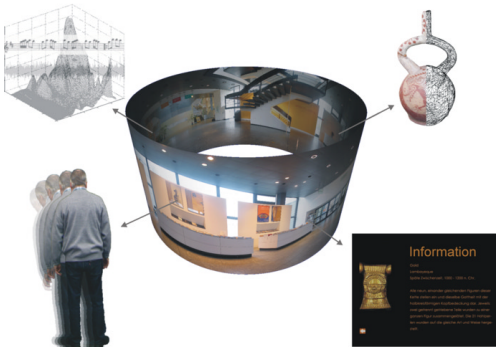

Figure 2: Scene elements of our MPEG-4 player. Besides the panorama, dynamic video objects, interactive buttons, 3-D models or spatial sound can be added to the environment.

fies the current location. The MPEG-4 system also ensures that only visible data is transmitted avoiding long downloads of the entire scene. Thus, large high quality environments can be created that enable the user to immerse into the virtual world.

Although the acquisition of large panoramas is quite simple in principle, in practice, the situation is often much more complex. For example, people, objects, or clouds in the scene may move while capturing the single images. As a result, the pictures do not fit to each other properly and ghost images appear. Moreover, capturing 360 degrees of a scene may impose high demands on the dynamic range of the camera. Especially in indoor scenes, extreme changes in intensity may occur between windows and the interior. We have therefore investigated algorithms for the removal of moving people and objects in order to simplify the stitching. Multiple views are captured at different time instants and the covered regions are warped from the same areas in other views. Capturing the scene with different shutter times enables an spatial adaptive adjustment of the dynamic range and to create panoramas also for scenes with extreme brightness changes.

The paper is organized as follows. First, the MPEG-4 framework is described that is responsible for view-dependent rendering and streaming of panoramas, videos, and 3-D objects. In Section 3.1 the determination of focal length and lens distortion is described which supports the accuracy of the stitching. The algorithm for the removal of objects is illustrated in Section 3.2. Section 3.3 finally describes the local adjustment of dynamic range and provides examples from real panoramas.

2 MPEG-4 System for Panorama Streaming and Rendering

The system for panorama rendering uses MPEG-4 technology which allows local display or interactive streaming of the virtual world over the internet. The scene is represented very efficiently using MPEG-4 BIFS [8] and rendered at the client using our MPEG-4 player [9]. The basic scene consists of a 3-D cylinder textured with a high resolution panoramic image as shown in Fig. 4. Other scene elements like 2-D images, video sequences, 3-D audio as well as interactive scene elements, like buttons or menus can easily be added. Since alpha masks can be provided to create arbitrarily shaped video objects, moving people or objects in motion can be added to the static scene creating more lively environments. Buttons allow to walk from on room to the next (Fig. 1) by requesting new BIFS descriptions or to display additional information.

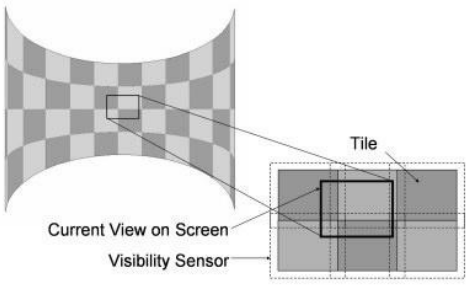

Figure 3: Subdivision of the panorama into small patches and visibility sensors.

Besides local display of the scene, MPEG-4 offers an interactive streaming technique, which transmits only data necessary to render the current view of the local user. We use the MPEG-4 client-server architecture based on the Delivery Multimedia Integration Framework (DMIF). DMIF allows to build applications unaware of the delivery technology details. For the particular panoramic scene with video sequences, 2-D image and 3-D audio objects, the movements of the pointer device is evaluated and the appropriate data for the desired view-

ing direction is requested from the server. In order to avoid streaming the entire panorama initially which would add severe delays, the high-resolution image is subdivided into several small patches. To each patch, a visibility sensor is added, which is active if the current patch is visible and inactive if it disappears again. Only active parts need to be streamed to the client unless they are already available there. The partitioning into patches and the visibility sensors are illustrated in Fig. 3. The visibility sensors are slightly bigger than the associated patch. This allows to prefetch the image patch before it becomes visible. The size of the sensors trade prefetching time with number of patches locally stored. This way, a standard compliant streaming system for panoramas with additional moving and interactive scene elements is realized.

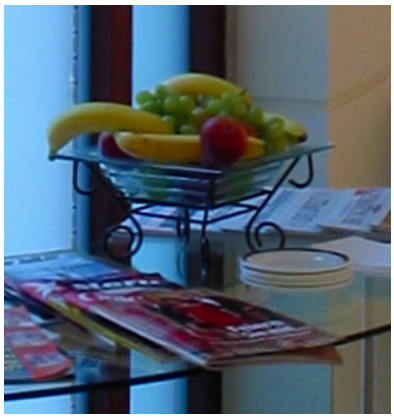

Figure 5: Closeup of the Adlon pool panorama. The content corresponds to the interior of the white box in Fig. 4.

3 Acquisition of Panoramas

The images for the panoramas are captured with a digital camera mounted on a tripod. For indoor environments, a wide angle lens converter is used to increase the viewing range. The camera on the tripod is rotated around the focal point by 15 to 30 degrees (depending on the viewing angle) between the individual shots. The resulting images are then stitched together into a single panorama using a commercially available tool (e.g., PanoramaFactory, www.panoramafactory.com). The output is a panoramic image as shown in Fig. 4, which is then subdivided into small patches of size 256x256 pixels for view-dependent streaming with the MPEG-4 system. With the current configuration, the resolution of the entire panorama is about 14000 by 2100 pixels which allows to view also small details by changing the zoom of the virtual camera. Fig. 5 shows a magnification of the white box in the panorama of Fig. 4.

3.1 Camera Calibration

In tests with several stitching tools, it has been evident that the accuracy of the results can be improved by determining focal length and lens distortions of the camera in advance rather than optimizing these parameters during stitching. We have therefore calibrated the camera with a model-based camera calibration technique [10]. The resulting intrinsic parameters like viewing angle and aspect ratio are passed to the stitching tool while the lens distortion parameters are used to correct the radial distortions in the images. Especially for the wide-angle lenses, severe distortions occur which have to be removed. Since the used camera can be controlled quite reproducible, it is sufficient to calibrate the camera once for various settings.

3.2 Object Removal

The stitching of multiple views to a single panorama requires the pictures to overlap in order to align them and to compensate for the distortions (due to projection, camera position, lenses, vignetting, etc.). After alignment, the images are blended to obtain a smooth transition from one image to the next. If a single camera is used and the images are captured one after the other, ghost images can occur in the blending area if objects or people move during capturing as, e.g., in the left to images of Fig. 6. These mismatches have to be removed prior to stitching.

In order to keep the number of images that have to be recorded low, we cut out the unwanted parts in the image by hand as shown in the third image of Fig. 6. The missing parts have now to be filled again. This can be accomplished either from the overlapping region of the previous view or from a second (or third) view that is recorded at a different time instant where the objects have moved again. In both cases the missing pixels must be warped from

Figure 4: Cylindrical panorama captured at the Adlon hotel, Berlin, Germany.

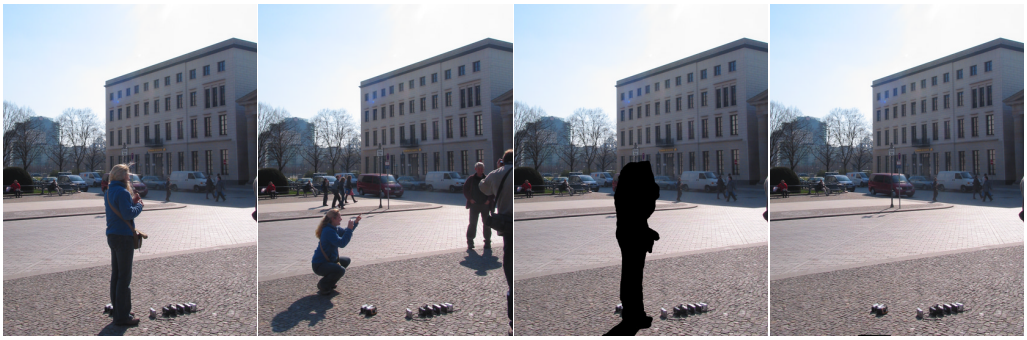

Figure 6: **Left images:** two pictures with a person that moves between the shots, **3rd image:** manually selected image mask, **Right:** final composed image ready for stitching.

the other view, filled into the region and blended with the background.

For the warping, we use the eight-parameter motion model

$$\begin{aligned} x' &= \frac{a_0 x + a_1 y + a_2}{a_6 x + a_7 y + 1} \\ y' &= \frac{a_3 x + a_4 y + a_5}{a_6 x + a_7 y + 1} \end{aligned} \quad (1)$$

that can describe the motion of a plane under perspective projection. For backward interpolation, x and y are the 2-D pixel coordinates in the reference frame while x' and y' are the corresponding coordinates of the previous frame or other source image. The eight parameters $a_0, \ldots, a_7$ describe the camera motion between the views.

If the motion is large, first feature points are searched, correspondences are established, and (1) is directly solved in a least squares sense. With the resulting parameters, the source image is roughly warped using (1) to obtain a first approximation.

This approximation is then refined using a gradient-based motion estimator [11]. Equation (1) is combined with the optical flow constraint equation

$$\frac{\partial I}{\partial x}(x' - x) + \frac{\partial I}{\partial y}(y' - y) = I - I', \quad (2)$$

which relates temporal with spatial intensity changes in the images. This equation is setup at each pixel position in an image area around the missing part to be filled. An over-determined set of linear equations is obtained that is solved in hierarchical framework. Since many pixels are used for the estimation, subpixel accuracy can be obtained. Again, the source image is warped according to the estimated motion parameter set and the missing pixels are filled in.

Figure 7: Part of the panorama of the Brandenburger Tor with people removed.

This warping is also done if multiple images are

captured from the same viewing position. Wind or vibrations can easily change the camera orientation slightly so that shifts of one or two pixels occur. The rightmost image of Fig. 6 shows the result of the warping and filling. The person in the left two images has been removed. In the same way, multiple people can be removed and Fig. 7 shows how the Brandenburger Tor in Berlin looks like without a crowd of people which can rarely be observed in reality.

The different acquisition time of the images can also lead to photometric changes, especially if clouds are moving. We therefore estimate also changes in color and brightness between the shots. A polynomial of second order

$$I' = c_0 + c_1 I + c_2 I^2 \tag{3}$$

is used to model a characteristic curve between the intensity value I of the reference frame and I' of the source image. Three unknown parameters c_0, c_1, c_2 are estimated for each color channel from the over-determined system of equations. Similar to the spatial warping, intensity changes can be corrected prior to filling of the missing image parts.

Fig. 8 shows some more examples for object removal by warping and illumination adjustment. The images are recorded in a tower at an airport where several people are working. During the capturing, several objects were moved and chairs were turned. The left side of Fig. 8 shows the images captured with the camera while the right images are corrected by warping from previous or succeeding views.

3.3 Dynamic Range Adaptation

The dynamic range in a scene can vary drastically which might lead to saturation effects in a camera capturing the scene. In 360° panoramas with a large number of possible viewing directions, the chance is high that there exist very bright and very dark regions. Especially in indoor scenes, drastic discontinuities can occur, e.g., at windows with a bright scene outside and a darker interior. Regular digital cameras are not able to capture such a dynamic range so that they often saturate at the lower or upper end.

These saturation effects can be avoided by combining multiple differently exposed images [12, 13]. In [14], it has been shown, that the simple summation of these images combines all their information due to the non-linear characteristic of the camera. In our experiments, the resulting panoramas, however, showed lower contrast, so we decided to use a locally adaptive summation similar to [12].

Figure 9: **Left:** one image from the Tower panorama. **Right:** Automatically computed mask to distinguish bright from dark image regions.

For each viewing direction, we capture three images. One with a long exposure time for dark areas, one with short exposure for bright regions, and one image that is located between the two. Then, a mask is computed that determines bright and dark areas in the image. For that purpose, the bright image is searched for saturated (bright) pixels and the dark one for saturation at the lower end. This information is combined to form the mask. Small regions in the mask are removed, morphological filters smooth contours, and an additional filtering add some blur in order to get smooth transitions between the different areas. Fig. 9 shows an example for the automatically computed mask and its corresponding image. Given the mask, a weighted sum of the images is computed, with the weights being locally determined by the image mask. Thus, the contrast remains high in bright as well as dark image regions.

This is illustrated in Fig. 10. The figure shows three differently exposed images from the interior of an airport tower with dark instruments in the foreground and a bright background. After image warping as described in Section 3.2 to account for moving objects, the images are adaptively combined into a new image shown on the lower right of Fig. 10 that reproduces the entire scene with high contrast.

4 Conclusions

A system for panoramic imaging based on MPEG-4 is presented. The use of the MPEG framework enables both streaming and local display of the scene. Moreover, interactive elements like buttons and menus or objects by means of videos, images, and 3-D computer graphics models can be added into the general BIFS scene description. This allows to enrich the static panorama by people or other dynamic objects as well as view-dependent audio in order to create a more realistic environment. We have shown that, e.g., moving people in the real scene and wide dynamic range of brightness can complicate the creation of panoramas. Algorithms have been presented to remove unwanted objects and to locally adjust the dynamic range, thus improving the quality of the high-resolution panoramas drastically.

5 Acknowledgment

The work presented in this paper has been developed with the support of the European Network of Excellence VISNET (IST Contract 506946).

References

[1] H.-Y. Shum and L.-W. He, "A review of image-based rendering techniques," in *Proc. Visual Computation and Image Processing (VCIP)*, Perth, Australia, June 2000, pp. 2–13.

[2] R. Szeliski and H-Y. Shum, "Creating full view panoramic image mosaics and environment maps," in *Proc. Computer Graphics (SIGGRAPH)*, 1997.

[3] S. E. Chen, "QuickTime VR - An image-based approach to virtual environment navigation," in *Proc. Computer Graphics (SIGGRAPH)*, Los Angeles, USA, Aug. 1995, pp. 29–38.

[4] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. Computer Graphics (SIGGRAPH)*, New Orleans, LA, USA, Aug. 1996, pp. 31–42.

[5] H.-Y. Shum and L.-W. He, "Rendering with concentric mosaics," in *Proc. Computer Graphics (SIGGRAPH)*, Los Angeles, USA, Aug. 1999, pp. 299–306.

[6] I. Feldmann, P. Eisert, and P. Kauff, "Extension of epipolar image analysis to circular camera movements," in *Proc. International Conference on Image Processing (ICIP)*, Barcelona, Spain, Sep. 2003.

[7] P. Eisert, "3-D geometry enhancement by contour optimization in turntable sequences," in *Proc. International Conference on Image Processing (ICIP)*, Singapore, Oct. 2004.

[8] *ISO/IEC 14496-1:2002, Coding of audio-visual objects: Part 1: Systems, Document N4848*, Mar. 2002.

[9] C. Grünheit, A. Smolic, and T. Wiegand, "Efficient representation and interactive streaming of high-resolution panoramic views," in *Proc. International Conference on Image Processing (ICIP)*, Rochester, USA, Sep. 2002.

[10] P. Eisert, "Model-based camera calibration using analysis by synthesis techniques," in *Proc. Vision, Modeling, and Visualization VMV'02*, Erlangen, Germany, Nov. 2002, pp. 307–314.

[11] P. Eisert, E. Steinbach, and B. Girod, "Automatic reconstruction of stationary 3-D objects from multiple uncalibrated camera views," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 2, pp. 261–277, Mar. 2000.

[12] S. Mann and R. Picard, "On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures," in *IS&T's 48th Annual Conference*, Washington, May 1995, pp. 422–428.

[13] P. E. Debevec and J. Malik, "Recovering high dynamic range radience maps from photographs," in *Proc. Computer Graphics (SIGGRAPH)*, 1997.

[14] M. D. Grossberg and S. K. Nayar, "High dynamic range from multiple images: Which exposures to combine," in *Proc. ICCV Workshop on Color and Photometric Methods in Computer Vision (CPMCV)*, Oct. 2003.

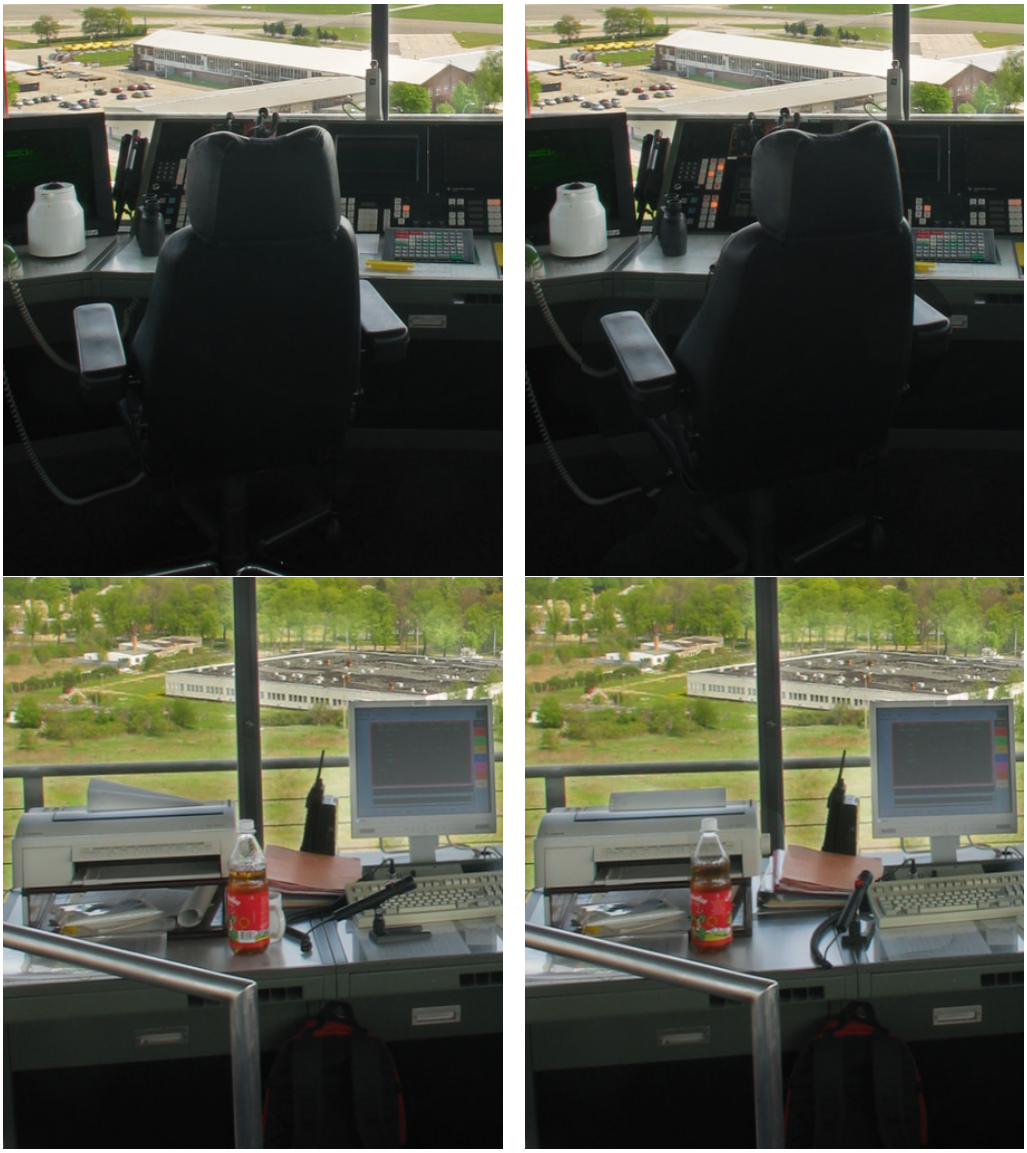

Figure 8: **Left:** original images from an airport tower. **Right:** same images after warping several objects from other views. The chair is rotated, printer, microphone, and bottle are moved appropriately.

Figure 10: **Upper row and lower left:** Differently exposed images from the interior of an airport tower. **Lower right:** combined image with high contrast in dark as well as bright regions.