Tomorrow's Digital Photography

Gerald Peter*
Vienna University of Technology

Figure 1: a) - e): A series of photograph with five different exposures. f) In the high dynamic range image generated from a) - e) mapped with a gradient-domain based operator by Fattal et al. [Fattal et al. 2002], all relevant contrast details in the scene are recognizable. g) Glare is an undesired and unavoidable effect at photos when the camera is pointed against a bright light source. h) This glare reduction approach from Raskar et al. [Raskar et al. 2008], based on the property that glare behaves like high frequency noise in 4D ray space, tries to statistically analyze the 4D ray space inside the camera, to classify and to remove the glare using a structured high frequency occlusion mask near the camera sensor. i-k) Refocussing after an image is taken, using a plenoptic camera [Ng et al. 2005]. l) A prototype of a plenoptic camera built by Ng et al. [Ng et al. 2005] allows refocussing.

Abstract

This article gives an overview about current research topics in digital photography. High dynamic range, high resolution panorama capturing, flash matting, scene completion and computational photography are features tomorrow's digital cameras will profit from. Image editing possibilities are strongly increasing after an image is taken. Users will be able to replace the background of a photo by taking a flash/no-flash photo pair, refocus images after they are taken and removing unavoidable glare effects in images with extended camera optics. Users will profit from large photo collections on the internet to interactively browse in 3D through automatically generated paths. Photocollections will also be used to find an automatically generated suitable and seamless replacement for a disturbing object of a photograph that was cut-away. Tomorrow's digital camera just as today's multimedia mobile phones might be able to establish an internet connection to feature these image-collection-based technologies.

1 Introduction

Photography is the process of creating still or moving pictures by recording radiation on a sensitive medium, such as film or electronic sensors. The idea of photography goes back to the 5th century B.C., when first Mo Ti described the pinhole camera. In the 20s of the 19th century, when the chemical photography was developed, the first permanent photograph was produced by the French inventor Nicphore Nipce, which required an eight-hour exposure. Based on experiments with silver compounds, Louis Daguerre took the first photo of a person in 1839 with an exposure time of several minutes. Between the 50s and the 80s of the 19th century photographic plates were common for capturing photographs. Finally the technology of film was developed in 1884 by George Eastman. Film replaced the photographic plates and remained the leading (analog) photographic technology until today. 1908 a method to reproduce colours photographically was introduced based on the phenomenon

*e-mail: gerald.peter@aon.at

of interference.

1981 the first consumer camera which needs no film, was introduced by Sony, which saved the images on disk and outputs them on television. But this camera was not fully digital. The first commercial fully digital available camera was unveiled by Kodak in 1990 which revolutionized photography. Digital cameras uses electronical image sensors and stores the captured images on a digital medium.

Today digital cameras are used in everyday life. The today's consumer standard camera is a digital camera which stores captured images on a digital medium, which can be easily watched, edited, stored on pc and sent over the internet. Compared to traditional analog photography, digital photos are very easy and nearly unlimited to reproduce.

Today's digital images are typically stored in RGB32 format, which allows to distinguish more than 16 million colours. Anyway, the contrast which can be captured is not enough to reproduce images with a match between the visual experience of the nature. In fact, with current output devices this match is not possible. But it is possible to extend the capabilities to use high dynamic range (HDR) imaging. HDR allows to record all relevant tones of a scene, even if very bright (e.g., direct sunlight) or very dark regions (e.g., shadows). Figure 1 (a) - (e) shows a scene with five different exposures, where either the interior or the outdoor scene is recognizeable. Figure 1 (l) shows that with HDR imaging all relevant contrast details in the interior as well as in the outdoor scene stay preserved. HDR imaging will be described in Section 2.

Increasing memory capacity of small digital media, allows to acquire more and more image data. The additonally available memory can be used for higher resolution, high dynamic range, higher depth of field, wider field of view, spectral information and the light ray distribution (i.e., a light field). Because today's consumer cameras are limited in possibilities of capturing, the trend of current research topics for the photography of tomorrow goes in the direction of processing several images of one scene to extract all these parameters. For example capturing several photos with different exposure time allows high dynamic range imaging. Capturing photos with different spectral filters allows to capture spectral information.

Images with high-resolution up to gigapixels, high dynamic range and higher field of view (panorama) can be only processed at one exposure with special unique camera devices. One approach which produces these images with the use of a today's consumer digital camera by processing a big amount of specific geometrically aligned images, is shown in Section 3.

Digital photography allows extensive post-processing by the user. For example extracting the background of the image manually and replacing it with another suitable background, such that the viewer can not recognize the post-processing can be done. A new method will be shown to automatically recognize a foreground object and the background. The extraction and removal of the background is made by taking a flash/no-flash image pair which will be shown in Section 4.

Computational and light field photography enhance or extend the capabilities of digital photography. The output of these techniques is an ordinary photograph, but one that could not have been taken by a traditional camera. In Section 5 digital cameras with extended optical systems acquiring additional information about the scene will be shown. The capturing from a 2D image gets replaced by the 4D light field acquisition. The 4D light field defines the directional ray distribution of light and can be captured for example with a light field camera which positions an lens array between the camera's main lens and the camera sensor or a programmable aperture introduced in Section 5. With the light field it is for example possible to compute the depth field for each pixel in the scene, to reduce glare effects (see Figure 1 (g) and (h)) and to refocus the image after it is taken (see Figure 1 (i) - (k)). Figure 1 (l) shows a prototype of a light field plenoptic camera, which looks like a traditional camera.

Large online digital photo collections like Flickr can be used for further "intelligent" processing. One approach shown in Section 6 tries to automatically find a seamless replacement of a cut image by scanning an image database with over two million photos categorized into several image classes. Online image collections are vast and unstructured. A new approach, also shown in Section 6 tries to extract the camera viewpoints of all photos by a feature matching algorithm and provides automatically generated controls to browse these photos in 3D with image-based rendering methods. Tomorrow's cameras might be able to establish a connection to the internet to implement these features based on large image collections.

2 High Dynamic Range Photography

The RGB Color System can represent only a limited range of colors and contrast. One option to extend the RGB color system is the XYZ color space. But both systems can only describe contrast relations up to two orders of magnitudes. Very dark and very bright parts of an image cannot be represented in one image with the RGB or XYZ color encoding. Human observer perceive details in scenes that span 4-5 orders of magnitude in luminance and adapt in minutes to over 9 orders of magnitude. High dynamic range images represent all details of contrast, which are perceivable by humans. HDR allows to capture contrasts from direct sunlight to shadows in one image. [Debevec et al. 2004]

A HDR captured image is shown in Figure 1 (f) which was recorded by 5 photos with varying exposure (see Figure 1 (a) - (e)).

The range of radiances recorded in each photo is limited, so not all details can be displayed at once. For example in Figure 1 (a) - (e), details of the room interior cannot be displayed at the same time as those of the bright outdoor scene. An algorithm is applied to the five images to recreate the high dynamic range radiance map of the original scene. The generated high dynamic range image is passed to a tone mapping operator, in this case a gradient domain operator by Fattal et al. [Fattal et al. 2002], which transforms the image into a low dynamic range image suitable for viewing on a monitor. As shown in Figure 1 (f), after tone mapping, all relevant constrast details in the room interior as well as in the outdoor scene are recognizeable.

HDR is not yet on the market, because RGB is the de facto output standard and the benefits of HDR without HDR displayable devices are poorly understood. There are some HDR monitors available, but they are very expensive and are not available on the consumer market. Viewing an HDR image on a common monitor requires tone mapping, which maps the HDR tones to the limited range of the output device. [Debevec et al. 2004]

2.1 Capturing HDR

Most cameras have not a built-in HDR function. HDR capturing is done by taking a set of several photos with varying exposure [Debevec et al. 2004] . The shutter speed and the aperture adjustment can be changed or neutral density filters can be stacked to vary the exposure. By varying the shutter speed, the exposure is modified directly. This usually is an accurate and repeatable process. Disadvantages are noise artifacts and blurring of moving objects while

long exposure times. Aperture variation can be used, if the photograph run out of shutter speed variation. This approach is not very recommended for HDR capturing because the aperture adjustment changes the depth of field and it has also a limited range of exposure variation. Another approach to reach a very wide range of exposure variation is the use of stacked neutral density filters. A problem that occurs with stacked neutral density filters is that they shift the image, and are rarely truly neutral.

To store high dynamic range images different possibilities and formats exist. LogLuv24, LogLuv32, RGBE, XYZE, EXR, scRGB, scRGB-nl, scYCC-nl are different standards for HDR image formats. OpenEXR is an open standard and has with 48 bits per pixel the highest accuracy and color fidelity of all formats. [Debevec et al. 2004]

2.2 Tone mapping

HDR cannot be displayed or printed with common output devices. The contrasts of an HDR image has to be remapped to the displayable contrast interval for viewing on a monitor or print out. As described in the talk by Weidlich [Weidlich 2008], the following categories of tone mapping approaches exist:

- Global methods
  - Linear scale factor
  - Non-linear scale factor
- Local methods
- Perceptual approaches
- Gradient domain operators

The global tone mapping algorithms use for all pixels of the hdr image the same computation (e.g. scaling). Figure 2 shows a histogram with the pixels of a hdr image. The red area shows the adjustable contrast interval in which tones are linear remapped. The borders and the size of the interval can be adjusted.

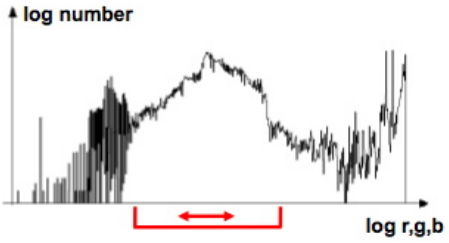

Figure 2: This histogram shows the pixels of a high dynamic range image. Simple tone mapping is done by a linear scale function, which maps tones from the HDR image of an adjustable contrast interval (red) to a low dynamic range image for displaying on a common output device, e.g. a monitor. [Weidlich 2008]

The results of linear mapping cannot compete the perceptual impression like non-linear methods. Better results are achieved by exponential mapping, because the exponential function correspond to the human perception. Figure 3 shows the exponential mapping function with the original contrast interval of the HDR image and the displayable contrast interval of the outputdevice.

The local tone mapping approach considers differences between various parts of the image. This method is similar to techniques

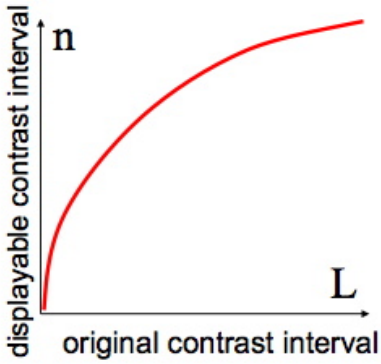

Figure 3: Exponential tone mapping, which maps the tones of a high dynamic range image to the displayable contrast interval of a common output device, e.g. a monitor. [Weidlich 2008]

from photography, where the image is separated into zones to determine the brightness of targets. A local kernel of variable size is used for the final tone reproduction step.

Perceptual tone mapping approaches use results from physiology and psychology in order to reproduce the behaviour of the human visual system. In a two-pronged perceptual approach, first by a person's impression of a scene is determined and secondly this sensation is approximated using a display device.

Gradient domain operators manipulates the gradient field of the high dynamic range image. The operator presented by Fattal et al. [Fattal et al. 2002] manipulates the gradient field of the luminance image by attenuating the magnitudes of large gradients. A low dynamic range image is obtained by solving a Poisson equation on the modified gradient field. This method enables drastic dynamic range compression, while preserving fine details and avoiding common artifacts. The result of tone mapping with this operator is shown in Figure 1 (f).

In the future maybe most of the commercial digital cameras will be able to capture a High Dynamic Range image automatically with one shot. The tones of a taken HDR image might be remapped through user interaction directly on the camera. HDR enables the photograph to capture more lighting detail, especially in complex lighting situations.

3 Giga Pixel Images

Taking images with resolutions up to giga pixels enables the user to explore all details. Several condentioned photos can be extracted from one gigapixel image. For example with one high resolution panorama of a city each building can be explored.

Capturing such a high resolution image with a single exposure needs special and relatively expensive equipment, currently not available on the consumer market.

Kopf et al. [Kopf et al. 2007] have demonstrated a system for the creation, processing, and interactive display of images with very high resolution, high dynamic range and wide angle fields of view, that combines hundreds of images captured with a conventional camera. Figure 4 shows three different zoom levels of a 1.5 giga pixel image.

Figure 4: Viewing a 1.5 giga pixel panorama image captured by Kopf et al. [Kopf et al. 2007] at three different zoom levels.

Figure 5: A device built by Kopf et al. [Kopf et al. 2007] which captures a high resolution, high dynamic range panorama image with a conventional camera . This device automatically adjusts angle and takes a shot in each iteration. After the capturing procedure, which records 250-800 images and takes between 30 and 90 minutes has finished, the images are stitched together with a radiometric and geometric alignment algorithm, which produces a high resolution, high dynamic range panorama image.

A specialized image capturing device allows them to acquire images with very high resolution. Efficient methods for the geometric alignment, radiometric alignment and tone-mapping automatically produce smooth, convincing results. The results can be displayed at interactive rates with according viewing software that smoothly adapts both the projection and tone mapping.

3.1 Capturing

Special cameras to record high resolution and high dynamic range images are not available on the consumer market and relatively expensive in comparison to normal cameras. The device Kopf et al. constructed shown in Figure 5, enables generation of these images with a normal camera. This device automatically adjusts angle and takes a shot in each iteration. One capturing procedure records 250-800 images and takes between 30 and 90 minutes. There is no need to capture the full dynamic range in each shot to cover a large dynamic range in the overall composite image. In this approach only the best exposure for one angle is taken.

Kopf et al. also developed an algorithm which construct from a vast number of individual shots one big gigapixel image. This algorithm handles both the scale and the high dynamic range issues. The only manual step of the geometric and radiometric alignment pipeline needed is to specify a neutral gray point in the scene. The rest of the processing is done automatically. Capturing images was done in the linear domain where the radiometric alignment and composition techniques from a seamless HDR image was constructed with no blending at all. In a further tone-mapping process the final result is produced.

3.2 Viewing

Kopf et al. developed a viewer for giga pixel panorama images which displays the current field of view (FOV) with a projection that best fits that FOV and smoothly varies the projection as the FOV changes. Further the tone mapping is adapted to the average luminance and contrast of the current image content. The user interface for the viewer contains controls for panning, which is mapped to mouse motion with the left button down, and zooming which is mapped to three different mouse actions. The pan position and zoom level defines the portion of the image to be displayed.

Images with a FOV up to 60 degrees can be viewed through a perspective projection on a monitor without visible distortions. If the FOV increases up to 80 degrees the distortions get visible. Another curved projection methods like spherical or cylindrical mapping can be used to ”unwrap” the image onto the flat monitor and decrease the distortion effect of perspective projection for wide angle views. The curved projections also incur distortions. However these distortions are less than perspective distortions for very wide angles.

The viewer from Kopf et al. provides in each zoom and any field of view setting the optimal projection by smoothly adapting the projection from perspective for small FOVs to curved for large FOVs. Figure 6 shows an example of zooming with perspective projection (top row), cylindrical projection (bottom row) and the adaptive method (center row). If it is zoomed out, the perspective projection exhibit strong distortions. Zooming in brings good results with perspective projection. The cylindrical projection produces undistorted images when zoomed out, but looks unnatural in the case of zooming in. The adaptive projection method described by Kopf et al. combines the advantages of both.

Smoothly interpolating between perspective and curved projections during panning and zooming is realized by bending, scaling, and rotating the projective surface within the world coordinate system. Interpolating between cylindrical and perspective projections is accomplished by unbending the projective surface from a cylinder to a plane. This can be viewed as increasing the radius of the cylinder while keeping the viewpoint at unit distance away from the cylinder wall.

The produced panoramas are initially stored in high dynamic range and require tone mapping to map the HDR image to the limited dynamic range of the monitor. The tone mapping operator which is used, combines a single (possibly manually-crafted) global tone mapping with a fast interactive local histogram-based tone mapping.

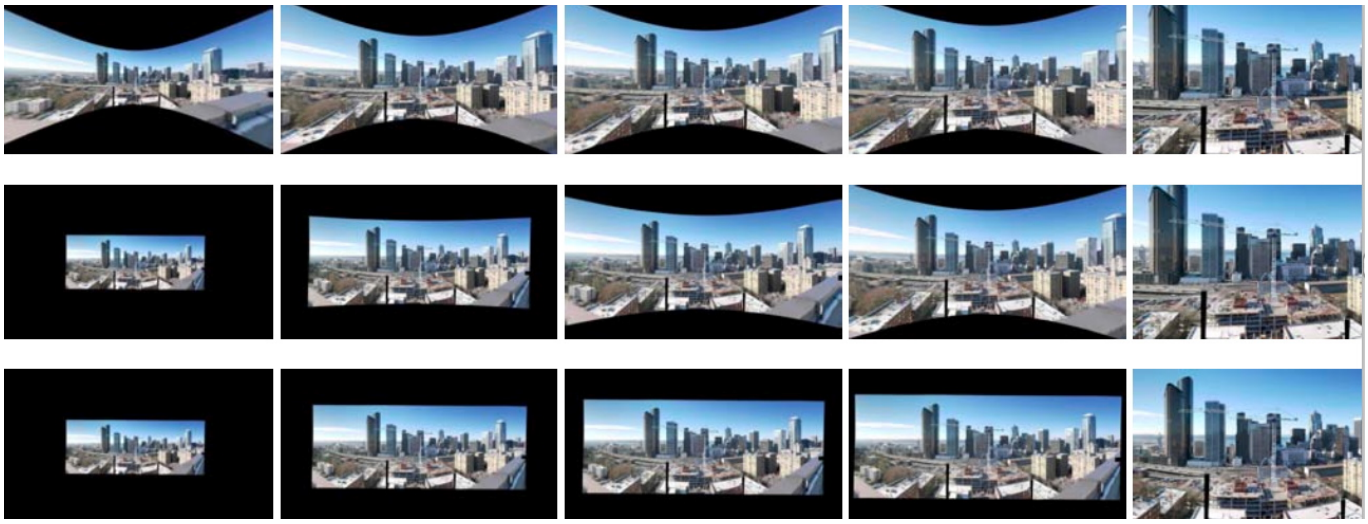

Figure 6: Zooming with perspective projection (top row), cylindrical projection(bottom row), and an adaptive projection method developed by Kopf et al. [Kopf et al. 2007] (center row). Perspective projection produce natural images at narrow field of views (right column), but generates strong distortions when zoomed out (left column). Cylindrical projection produces undistorted results when zoomed out (left column), but generates unnatural results in narrow field of views (right column), because straight lines appear curved. The adaptive projection method developed by Kopf et al. [Kopf et al. 2007] adapts between these two projections to provide a natural appearance in each arbitrary zoom level. In wide field of views the cylindrical projection and in narrow field of views the perspective projection is applied.

4 Flash Matting

Flash matting separates a foreground object from the background by taking a flash/no-flash image pair. This method, introduced by Sun et al. [Sun et al. 2006] needs no user interaction (except the selection of the new background image) and could be implemented on tomorrow's camera. An advantage of this method compared to previous matting approaches is, that it needs no special equipment and handles scenes with complex foregrounds very well, also when background and foreground have similar colors.

4.1 The matting problem

Matting in computer graphics handles the problem to separate a foreground object from the background. The problem can be abstracted through following equation:

$$I = \alpha F + (1 - \alpha)B \tag{1}$$

This equation is called the compositing equation where the image I is composed by the foreground object F, the background B and the α matte. The goal for a given image I is to estimate F, B and α. Fractional opacities (between 0 and 1) are important for transparency and motion blurring of the foreground element, as well as for partial coverage of a background pixel around the foreground objects boundary [Chuang et al. 2001].

4.2 Matting Methods

There are two trivial methods to extract the foreground object. In blue screen matting the background B consists of one color which is known. This greatly simplifies the matting problem. In difference matting the background B is known, which do not have to be a constant color. In difference matting one photo with the foreground and one photo without the foreground photo is made, which results by evaluating the difference image in the foreground object F.

In natural image matting the matte is computed from one single image without background information. As described in the article by Sun et al. [Sun et al. 2006], first, the input image is manually partitioned into definitely foreground, definitely background and unknown regions. These three regions defining the trimap. Then, α, B and F are estimated for all pixels in the unknown region. In natural image matting there are both statistical sampling based approaches as well as gradient domain based approaches for the automatical generation of mattes. To date, given a good trimap, Bayesian matting is still regarded as one of the most robust methods for automatically generating mattes. The major weaknesses of this methods is to resolve ambiguity when the foreground is similar to the background and to handle complex foregrounds and backgrounds.

Bayesian matting proposed by Chuang et al. [Chuang et al. 2001] tries to extracting a foreground element from the background image by estimating an opacity for each pixel or the foreground element. It models both the foreground and the background color with spatially-varying sets of Gaussians, and assumes a fractional blending of the foreground and background colors to produce the final output. It uses a maximum-likelihood criterion to estimate the optimal opacity, foreground, and background simultaneously. Chuang et al. claim that their algorithm handles objects with intricate boundaries, such as hair strands and fur better than previous techniques.

Flash matting described by Sun et al. [Sun et al. 2006], which will be shown in the next subsection, extends Bayesian matting algorithm, does not need user interaction. Flash matting produces superior results than Bayesian matting. and uses the and delivers even better results than Bayesian matting. Due to the lack of user interaction it could be built into future digital cameras.

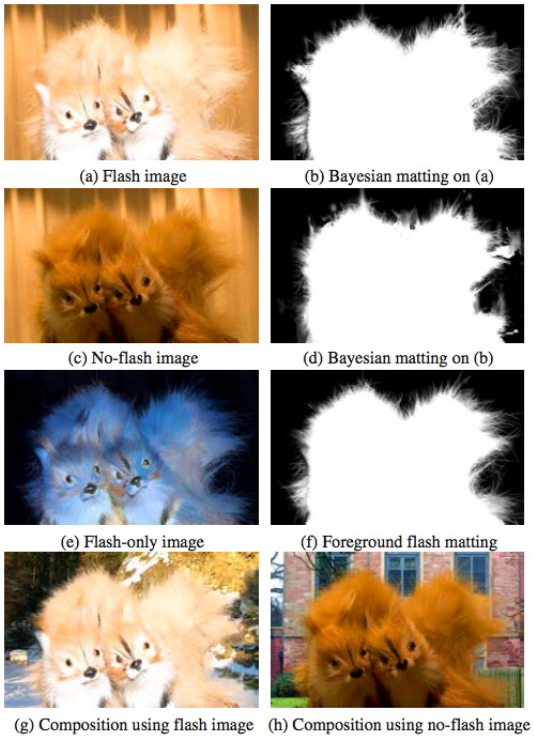

Figure 7: An example where foreground flash matting, proposed by Sun et al. [Sun et al. 2006] was used and compared with Bayesian matting. Flash matting is a novel method for extracting a foreground object from a background object by taking a flash/no-flash image pair. (a) shows the flash image. (b) shows the Bayesian matting result for (a). (c) shows the no-flash image. (d) shows the Bayesian matting result for (c). As seen in (d) Bayesian matting has problems concerning the separation of regions where the fore- and background have similar colors, which leads to artifacts. (e) shows the difference image between the flash and the no-flash image, which is denoted as flash-only image. (f) shows the foreground flash matting result which leads to a better seperation result than Bayesian matting which is shown in (b) and (d). (g) shows a composition with a new background image using (a) and (f). (h) shows another composition with a new background image using (c) and (f)

4.2.1 Flash matting

Sun et al. [Sun et al. 2006] presented an algorithm for flash matting which is able to separate the foreground image to the background by taking a flash/no-flash photo pair of a static scene. In this approach no special studio equipment is needed and both the trimap and the matting result is generated automatically. The algorithm is applicable for any kind of scenes with an unknown and maybe complex background even when the back- and foreground colors are similar. The *foreground flash matting* problem, which is shortly described in this section, is the straight forward method for flash matting without optimization.

Two assumptions are made:

- Only the foreground changes dramatically when the photo is made with flash. This assumption is justified because of the decrease of flash intensity with the inverse of the squared distance. A distant background is poorly lit by the flash compared to the foreground.
- The input image pair is pixel aligned. This algorithm only works for static scenes with a foreground which is lit from the flash and a background which is not influenced by the flash.

With this assumptions the difference of the flash/no-flash photo pair (see Figure 7 (a) and (c)), which is defined as flash-only image, eliminates the background (see Figure 7 (e)). However the foreground flash matting problem is not solved. A trimap is still necessary for further processing. A trimap is automatically generated from the flash-only image by a method similar to Canny's two-pass method for edge detection. For further details, how this algorithm works see [Sun et al. 2006]. To get a high quality matte, at first the α matte is computed from the flash-only image with the Bayesian matting algorithm using the automatically generated tripmap (see Figure 7 (f)). Afterwards the recovered α is used to extract the foreground F either from the flash or no-flash image (see Figure 7 (g) and (h)). Figure 7 (b) and (d) show the flash mattes, which are generated only from either the flash or no flash image, which clearly brings worse results in matting.

The straight forward foreground flash matting approach, which performs better than the single image Bayesian matting approach, is poor conditioned if the absolute pixel value of the flash-only image gets very small, or the foreground has a low reflectivity, or the surface normal is nearly perpendicular to the flash direction. Better results are achieved with the *joint Bayesian flash matting*. Joint Bayesian flash matting handles unknown pixels with statistical methods which leads to better results in poor conditioned pixel areas.

Another problem, which occurs generally in flash photography, is *self-shadowing*, which is caused by depth discontinuities within the foreground object and significant displacement between the flash unit and the camera's optical center. Shadow pixels cause a value of zero in the pixel of the difference image which leads to false results. These errors are small and can be reduced by joint Bayesian flash matting. To avoid self-shadowing at objects with large internal depth discontinuities it is necessary to use a ring-flash.

The capturing of the flash/no-flash image pair can be done with the continuous shot mode of current digital cameras, where up to 5 images per second can be taken, where the first image is shot with the flash light.

5 Light Field Photography

Future photography will not deliver only a 2D image of the 3D world. Light field photography captures light fields which consist directional information of the light rays of the scene. These light fields can be used for different applications, which will be shortly shown in this section.

As described in an article by Zyga [Zyga 2007], Dave Story from Adobe presented a camera with an array of 19 special lenses, where at one shot multiple angles/views are captured. With a special algorithm it is possible, to determine depth information for every pixel in the scene. It is for example possible to simply erase everything in the image, that appears at or beyond a certain distance (comparable to Flash matting as described in Section 4). With light field photography it is also possible to re-adjust the focus or to view photos from different angles after they are taken.

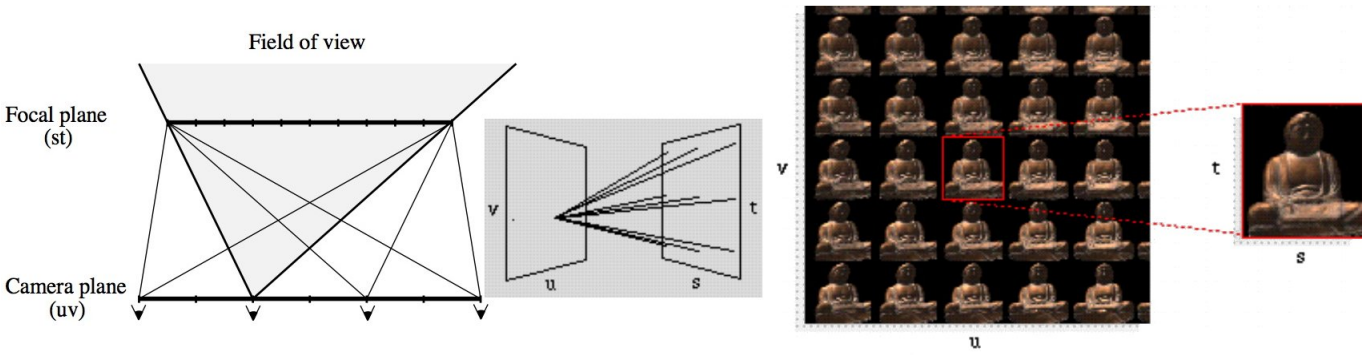

Figure 8: Left: A 4D light slab, which is defined as the beam of light entering one quadrilateral and exiting another quadrilateral, can be generated by taking a 2D array of images, where each image represents a slice of the 4D light slab at a fixed uv value and is formed by placing the center of projection of the camera at the sample location on the uv plane. The viewing geometry shown is used to create a light slab from an array images. Center and Right: The resulting 4D light field can be interpreted as an uv array of st images. Each image in the array (right) represents the rays arriving at one point on the uv plane from all points on the st plane (center) [Levoy and Hanrahan 1996].

5.1 Light Field Theory

Light fields were introduced in field of computer graphics in 1996 by Levoy and Hanrahan [Levoy and Hanrahan 1996]. The basic idea is to capture and process information of light rays. 10 years later Levoy [Levoy 2006] describes in his article the possibilities with full captured light fields which are:

- To fly around scenes without creating 3D models of them.
- To relight scenes without knowing the surface properties.
- To refocus photographs after they have been captured.
- To create unusual perspective or multi-perspective panorama views.
- To build 3D models of scenes.

Light fields opens new ways in photography, which are not possible in traditional photography. Some of these points will be shown in this section. First at all the limits of traditional photography optics and the basics of light fields will be shown.

The 4D light field Traditional photography projects one view of a 3D scene on a 2D image plane. At each grid point on the image, the integration of all light rays from all directions arrived, is measured and represented through a pixel color value. The information about the directional radiation of each light ray in the scene is lost in 2D photography through the implicit integration step.

As described in the article of Levoy [Levoy 2006], the 4D light field was first defined by Arun Gershun in 1936 and not only defines the pixels of the 2D projection of a 3D scene. It defines the amount of light denoted by L traveling along a ray in every direction through every point in space. The *plenoptic function* describes the radiance along all such rays in a region of 3D space illuminated by an unchanging arrangement of lights. It is a 5D function L(x,y,z, θ,ϕ) which is defined on each position (x,y,z) in space in all directions (θ,ϕ) and measured in watt per steradian(measures solid angle) per m^2. Current devices are not able to measure the light transport between two points of a concave objects. When all objects in the scene are restricted to their convex hulls, the definition of the 5D function implies redundant information because of the property that the radiance between two points remains constant along a ray. The 4D light field L(u,v,s,t) describes the radiance along rays in empty space and is with the convex object restriction equivalent to the 5D function.

As described in the article of Levoy and and Hanrahan[Levoy and Hanrahan 1996], the parameterization (u,v,s,t) defines lines of the 4D light field and their intersections with two planes. By convention, the coordinate system on the first plane is (u,v) and on the second plane is (s,t). An oriented line is defined by connecting a point on the uv plane to a point on the st plane, which is also shown in the center image of Figure 8, where several lines going through one point of the uv plane intersect the st plane. This representation is denoted as light slab, which represents the beam of light entering one quadrilateral and exiting another quadrilateral. The creation of a light slab is shown in Figure 8 on the left. As shown, a way to generate light fields is to assemble a collection of images. Figure 8 (right) shows a visualization of a light field where each image in the array represents the rays arriving at one point on the uv plane from all points on the st plane.

As described, the light field can be interpreted as two dimensional array of 2D images. Such a light field can be captured for example with a camera array or a plenoptic camera with a build-in lens array.

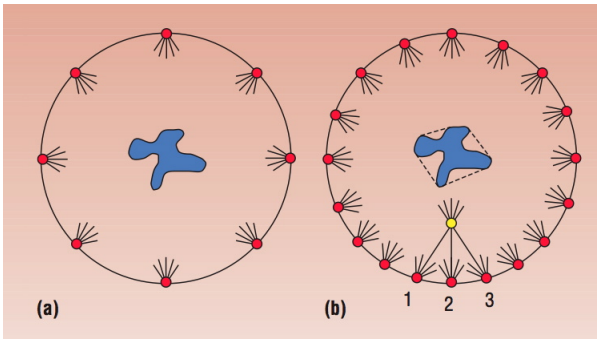

Figure 9: (a) Cameras are arranged along an arc pointing to the centered object. As long as the sphere is large enough to not intersect the objects convex hull, the collection of images is a 4D light field. This allows to fly around the object (blue shape) by flipping among closely spaced photographs of it (red dots). (b) Light field rendering: If the dots are spaced closely enough,the user can re-sort the pixels to create new perspective views without having stood there (yellow dot). [Levoy 2006]

Figure 9 schematically shows a simple case of a light field capturing. Many cameras are arranged along a circle pointing to the object in the circle's center. In each position one photo is captured. The

captured light field allows to fly around and towards the object by computing new views.

Image Reconstruction As described in the article of Levoy and and Hanrahan[Levoy and Hanrahan 1996], a completely captured light field of a scene allows all effects mentioned above as well as *image-based rendering*. Image-based rendering allows to render a scene from arbitrary viewpoint without an available 3D model. Instead it provides for each viewpoint a image of the scene. Advantages of this rendering approaches are, that the cost of interactively scene view does not depend on the scene complexity and that the image source can be taken from real photograph as well as from rendered models. However it is difficult to acquire a full light field of the scene.

All views of an object, which would allow image-based rendering could be generated from one light slab only if its set of lines include all lines intersecting the convex hull of the object. Unfortunately, this is not possible. Therefore, it takes multiple light slabs to represent all possible views of an object. Light field cameras are not able to capture multiple light slabs and full light fields of the scene because of the relatively small range of views, which does not enable 3D image-based renderings. As described in the article of Levoy [Levoy 2006], a special construction is built by the Stanford Computer Graphics Laboratory to capture multiple light slaps and as match as possible of the light field of a small object. This construction measures the light rays from many viewpoints around the object which has a camera mounted on a computer-controlled motorised arm.

Figure 9 shows schematically a light field capturing example, where several cameras are arranged along an arc pointing to a centered object. This allows to recompute arbitrary viewpoints of the object outside its convex hull. In Figure 9 (b) the yellow point represents a new generated view computed from the views 1,2 3. The central pixel of the "yellow view" is identical to the central pixel of view 2. The rightmost pixel in the "yellow view" equals to a pixel of view 1. If the number of captured photos from different viewpoints is large, perspectively correct views can be reconstructed. Interpolation of nearby pixels is needed to reconstruct the plenoptic function at any viewpoint outside the convex hull of the object. If there are not enough captures of different viewpoints, the renderings will contain ghosts arising from blending different views of an object. Light fields renderings enabling free-fly views around an object in 3D needs at least 1000 images. But the number of necessary captures is unlimited, so there also exists a light field capture of the Michelangelo's Night in Florence's Medici Chapel with 24.000 images.

Devices like the plenoptic camera or the programmable aperture camera captures the scene from a limited number of positions and angles which disables recomputing viewpoints like the example above, flying around a object. These devices capture the light field over a certain range of angles which limits the viewpoint displacements possibilities.

5.2 Methods of Capturing Light Fields

There are different possibilities for acquiring a light field. A complete 4D light field contains most visual information of a scene and allows various photographic effects to be generated in a physically correct way. There are several methods to acquire a light field:

- Moving camera: A simple method to capture the light field is to move a single camera and capture at each relevant position an image and the position and exposure information [Levoy 2006]. This method is restricted to static scenes.
- Arrays of cameras: A complex and inconvenient method for fast acquisition of a light field is to arranging several cameras in a specific geometric alignment. Wilburn et al. [Wilburn et al. 2005] built such an array with 100 cameras aligned on a plane.
- Plenoptic camera: Ng et al. [Ng et al. 2005] built a plenoptic camera which has a built-in microlens array, placed at the original image plane inside the camera. This method records the angular distribution of the light rays. The problem of this type of acquisition is the relatively low spatial resolution for reaching a useable angular resolution. It is very difficult to reach both, high spatial and angular resolutions. This camera enables refocussing (see Figure 1 (i)-(k)) after an image has been taken and looks like a conventional camera (see Figure 1 (l)).
- Programmable aperture: Liang et al. [Liang et al. 2008] extended a normal camera with a programmable aperture which places programmable non-refractive masks at the aperture of a camera provides a full sensor spatial resolution and enables to capture the light field through sequential multiple exposure without any additional optic elements and without moving the camera.

In light field acquisition the spatial resolution is the conventional pixel resolution of a camera. The angular resolution is the amount of directional information, that can be captured by a light field device.

5.3 Arrays of Cameras

An array of cameras can be used to capture a light field. Wilburn et al. [Wilburn et al. 2005] built a unique array of 100 cameras with capabilities of a system that would be inexpensive to produce in the future. The applications of this system include approximating a conventional single center of projection video camera with high performance along one or more axes, such as resolution, dynamic range, frame rate, and/or large aperture, and using multiple cameras to approximate a video camera with a large synthetic aperture. This allows capturing a video light eld, to which spatio-temporal view interpolation algorithms can be applied in order to digitally simulate time dilation and camera motion. The creation of video sequences using custom non-uniform synthetic apertures is also possible with this system.

5.4 Plenoptic Camera

The first prototype of a plenoptic camera was built by Ng et al. [Ng et al. 2005]. This camera looks like a conventional camera (see Figure 1 (l)), but has a built-in microlens array between the sensor and the main lens. This camera is able to sample the 4D light field on a single exposure. An analogy to the biology would be to take a human eye and replacing its retina with an insect eye (analogous to the microlens array). This camera enables refocussing after an image is taken, which is shown in Figure 1 (i)-(k).

One microlens does not measure the total amount of light at each location of the sensor. The microlens measures how much light arrives along each ray, in other words the directional lighting distribution at each location of the sensor.

The plane of the microlens array can also be interpreted as a sensor plane where at each grid point an array of ray-direction-dependent pixels are stored.

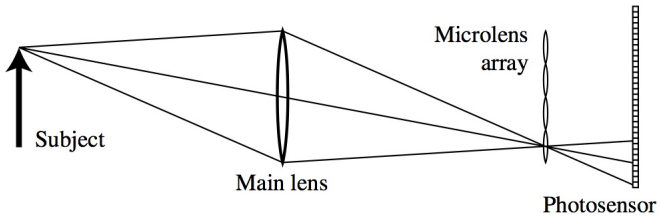

Figure 10: Optics of a plenoptic camera. This camera has a microlens array at the position where a normal camera has its sensor. The microlens array separates the converging rays into an image on the photosensor behind it. [Ng et al. 2005]

Figure 10 shows the optics of a plenoptic camera. Rays of light from a single point converge to a single point on the focal plane of the microlens array. Microlens at that location separates these rays of light based on direction, creating a focused image of the aperture of the main lens on the array of pixels underneath microlens.

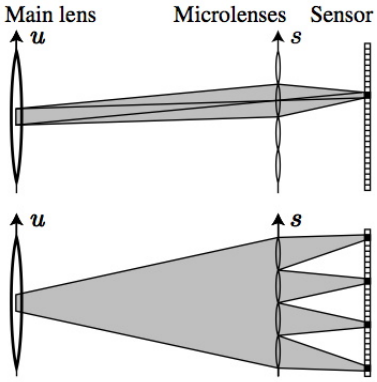

Figure 11: Top: The light that passes through a pixel passes through its parent microlens and through its conjugate square (sub-aperture) on the main lens. Bottom: All rays passing through the sub-aperture are focused through corresponding pixels under different microlenses. These pixels form the photograph seen through this sub-aperture. [Ng et al. 2005]

Aperture sizes (f-stops) of the main lens and the microlenses have to be balanced to avoid the rays from different microlens overlap on the photosensor if the main aperture size is too large or the sensor resolution is wasted in case the main aperture size is too low.

Figure 11 shows, that all rays passing through certain pixels, are going through the same sub-aperture on the main lens. Each pixel on every associated sensor area of a microlens belongs to a certain area on the aperture. That means that every ray passing through the aperture, is deposited in one of these pixels.

To form a photograph certain pixels of every microlens sensor area are extracted and composed to an output image. At the bottom of Figure 11 only one pixel (marked as black) of each microlens sensor is used to recompute an image with less aperture size.

As seen in the example, the aperture size of an image can easily be varied after capturing by changing the number of pixels of each microlens sensor area contributing to the final image. That means that the depth of field can be varied after capturing easily.

At the plenoptic camera (or also programmable aperture) approach the viewpoint can be varied by composing the image only with that rays going trough a small area at a certain position along the main lens. That means, if a light field photograph with a big aperture size is taken, it is possible to construct all sub-aperture perspectives. The disadvantage of a sub aperture image at plenoptic cameras is its low resolution in relation to the full sensor resolution. The resolution of the sub-aperture, where the image is constructed with one pixel per microlens, is restricted to the number of lenses. That generally means, that if the image of the main lens under the microlense is N pixels across, then the width of the sub-aperture is N times smaller than the width of the lens' original aperture. The programmable aperture method do not have this problem, but another disadvantage of this method is the sequential capture, which limits it to static scenes.

Refocussing is done with the plenoptic camera by summation of shifted versions of the images, that form through pinholes over the entire plane of the main lens. So only shifting and adding of sub-aperture images is necessary for refocussing.

Figure 12 shows the principle of the refocusing procedure with a plenoptic camera. Using a plenoptic camera, the plane of the microlens array can be interpreted a sensor plane where at each grid point an array pixels are stored for each ray direction. In the top image the middle yellow part of the rectangle object is in focus. The rays hit each other in one point on the focussed object as well as on the (virtual) sensor plane (which is in case of the plenoptic camera a microlens array). The top right image shows the image of one microlens. Each pixel corresponds to the light of one ray direction. On the top left of Figure 12 one example image is shown where the foreground woman is in focus and the people in background are not. To re-adjust the focus, in conventional photography the distance between lens and sensor plane gets varied. This step can be simulated in light field photography after exposure. In the bottom of the example a new focus is generated by constructing a virtual sensor plane which with decreased distance to the main lens. Each pixel of the virtual plane is computed through summation of certain rays of the acquired light field. The right bottom picture in Figure 12 shows three microlens images. Setting the focus plane behind the rectangle object, the rays which hit together at the camera's virtual focal plane are added together by taking the sum of certain pixels of the microlens images. The bottom left image shows the result of refocusing. Now the people in background are in focus.

In summary a plenoptic camera allows to create a light field with one capture using a built-in microlens array which captures the directional light distribution in each pixel. Refocussing and observer moving has been tested with this prototype. The range of the available viewpoints is limited by the diameter of the camera's aperture. In the test the plenoptic camera reached a focal ratio of f/22 after refocussing with f/4 capturing. The disadvantage of the plenoptic camera is the low resolution caused by the microlens resolution (300x300).

5.5 Programmable Aperture

The use of a programmable aperture which places programmable non-refractive masks at the aperture of a camera provides a full sensor spatial resolution to capture the 4D light field through sequential multiple exposure without any additional optic elements and without moving the camera [Liang et al. 2008].

Liang et al. [Liang et al. 2008] demonstrate a novel optimal

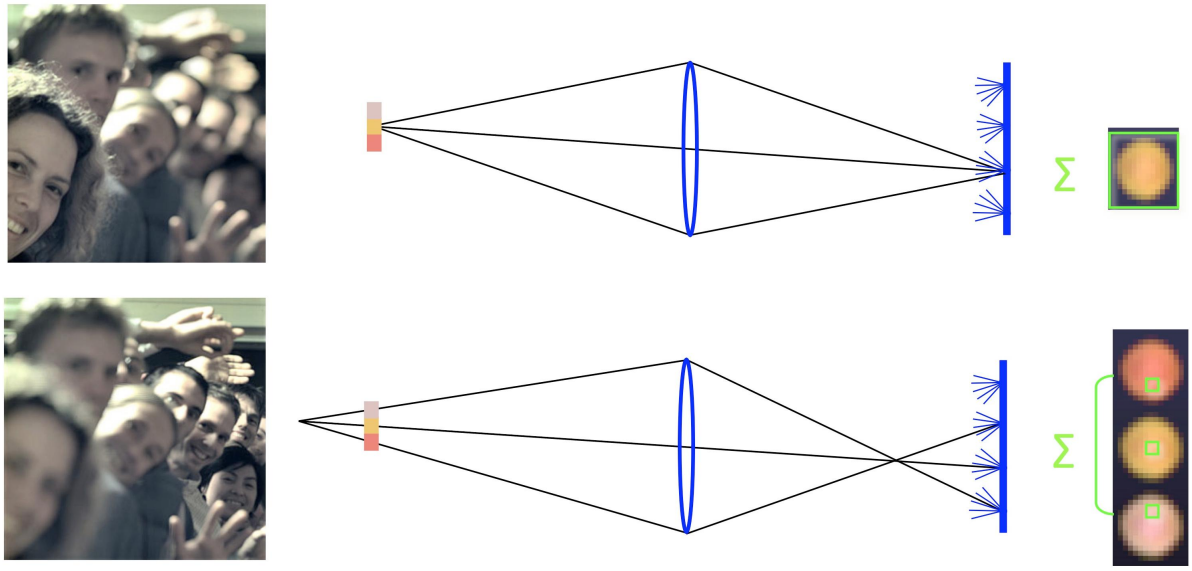

Figure 12: Principle of digital refocussing. The focus with a conventional camera is adjusted by variation of the distance between the main lens and the image plane. This is done virtually with a light field camera, which records the directional light distribution in each point of the image plane. Top left: Image captured with a plenoptic camera built by Ng et al.[Ng et al. 2005], where the people in are in focus. Top center: The associated illustration where a nearby object is in focus. The yellow region of the three-colored object is focussed, because one point of the object converge to one point on the image plane. The light of each ray corresponds to one pixel of the top right image. Top right image: Represents the directional light distribution of one point on the image plane. Each pixel represents the amount of light coming from one certain direction. A conventional camera would sum the light coming from all directions hitting the point on the image plane. The light field camera preserves the information, how much light comes from each direction. Bottom left: Image captured with a plenoptic camera built by Ng et al.[Ng et al. 2005], where the background was focussed after the image was taken. Bottom center: The associated illustration where a further afar point is focussed. To change the focus after the image is taken the distance between the main aperture and the image plane is virtually displaced by summing certain rays which converge to a point lying closer to the main aperture. Bottom right: Three different pixels which correspond to the light rays in the center bottom example are added together to compute one pixel on the virtual image plane which lies closer to the main aperture. The refocussed image is obtained by doing this step for all pixels on the virtual image plane. (Center and right illustrations taken from Levoy [Levoy 2005])

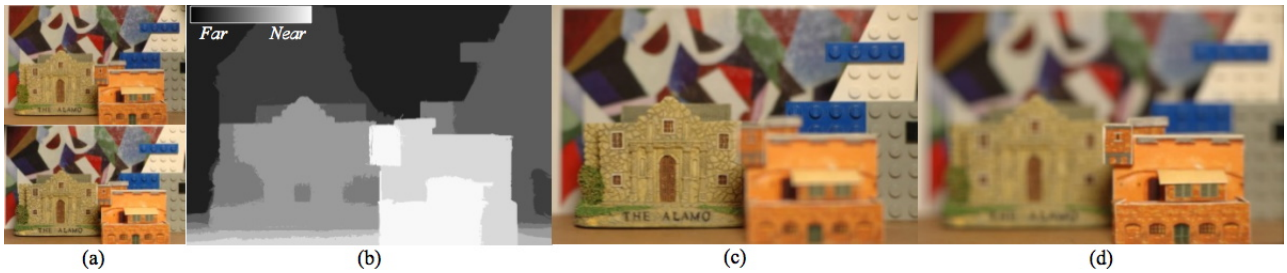

Figure 13: Capturing a lightfield and re-focus the image with the programmable aperture approach by Liang et al. [Liang et al. 2008]. (a) shows two demultiplexed light fields. (b) shows the computed depth field. (c) and (d) demonstrates re-focusing computed by the light- and depth-field

multiplex algorithm and two associated post-processing algorithms for the programmable aperture approach. Figure 13 (a) shows two input light field images, (b) shows the estimated depth map, (c) and (d) refocused images generated from the light field and the depth maps. The taken light field image has a resolution of $4 \times 4 \times 3039 \times 2014$. 3039×2014 is the spatial, and 4×4 is the angular resolution of the light field. That means that the aperture is partitioned into a 4×4 matrix, where each single region is sequentially captured. Figure 14 shows the prototype of this technique. An electronical programmable liquid crystal array which is put on a aperture of a single-lens reflex (SLR) camera where each region of the array can be turned on or off.

In traditional photography, a sensor integrates the radiances along rays from all directions into an irradiance sample and thus loses all angular information of the light rays. The goal of this method is to capture the light field which contains both the spatial and the angular information. The light field is captured sequentially such that for each sub-aperture the complete sensor image is captured (see Figure 15 (a)). For example, at first only the sub-aperture at the programmable aperture raster at one certain position is captured. That means, that only light rays can pass through this sub-aperture, which delivers for each sub-aperture image a directional information of the captured image. Then the image for the next sub-aperture is captured, until all sub-aperture images are captured. After all

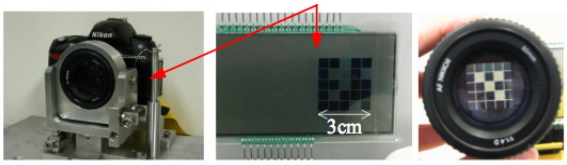

Figure 14: The prototype of the programmable aperture is realized with a programmable liquid crystal array to apply various patterns for multiplexing. This array is put onto the aperture of a conventional camera. [Liang et al. 2008]

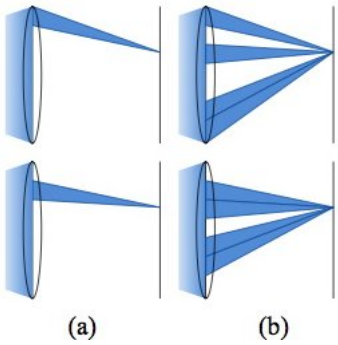

Figure 15: Light field capture using the programmable aperture method by Liang et al. [Liang et al. 2008]. (a) The programmable aperture method captures sequentially single samples of the aperture not to loose the angular information. (b) The problem of noisy results caused from the low light energy passing through the small area of the aperture in (a) is solved through multiplexing the aperture. Demultiplexing is done in a post processing step which leads to a quality improvement.

sub-aperture images are taken, each pixel of each sub-aperture image is known and the light field capturing process is finished.

There are two problems with this process. At first, the capture of the light field works sequentially and is as a consequence restricted to static scenes. In traditional cameras, light passes the whole aperture which decreases depth of field, but increases the exposure. In this approach only a small aperture and low exposure time is available which leads the bad exposure and a dark and noisy image. This defines the second problem which concerns noise artefacts, caused by the low light energy passed through a small sub-aperture.

To solve the second problem of noise artefacts, Liang et al. [Liang et al. 2008] developed a multiplex algorithm, which uses sub-aperture patterns to increase the light energy which is captured at each shot. If the aperture raster has a resolution of N as well as N different captured sub-aperture patterns, it is possible to reconstruct the sub-aperture image for each aperture raster element. Figure 15 (a) shows schematically the capturing process for each single sub-aperture which leads to noise artifacts caused by bad exposure. Each single sub-aperture image can be reconstructed from N multiplex pattern images as shown in Figure 15 (b).

The process is shown in Figure 13, where two demultiplexed light field images captured (a) with the multiplexed programmable aperture method. Figure 13 (b) shows the depth field estimated from the captured light field. Figure 13 (c) and (d) shows a refocussing example using the depth map. Liang et al. [Liang et al. 2008] show their approach in the programmable aperture method.

5.6 Reducing Glare

Glare is an undesired, unavoidable effects at photos when the camera is pointed against a bright light source, for example the sun. It arises due to multiple scattering and reflections of light inside the camera's body and lens optics and reduces image contrast, which causes fog and ghost effects.

There are two methods for removing these effects. Deconvolution by measuring a glare spread function is one of them to remove glare in a post processing step operating on the 2D image. A new and more effective method proposed by Raskar et al. [Raskar et al. 2008] removes the effect with computational photography approaches. The approach from Raskar et al., based on the property that glare is a 4D ray space phenomenon, tries to statistically analyze the 4D ray space inside the camera, to classify and to remove the glare. In ray space glare behaves like high frequency noise. An example of glare reduction is shown in Figure 1 (g) and (h), where Figure 1 (g) is the original and Figure 1 (h) the glare reduced image using the approach by Raskar et al. .

A traditional plenoptic camera has to compromise with the spatial solution according to the microlens array. The approach by Raskar et al. does not need the spatial structure of the ray-space and uses a structured high frequency occlusion mask near the camera sensor. It separates spurious rays in ray space.

There are different types of glare that have to be considered. Reflection glare appears as parasitic image when a strong light source causes a complex series of reflections among the lens surfaces. Both ghosts and flares are effects arising from reflection glare. Ghosts appear as aperture-shaped reflections in a position symmetrically opposite the light source and a flare as more uniform fogging of a large image area. Scattering glare is the second type of glare, which arises from diffusion at the lenses. The optical element act as mild diffusers. Reflection glare and scattering glare overlaps in a 2D image and is difficult to automatically identify, classify and remove. In 4D ray space the distinction between both types of glare is clearer. Adding a mask to the camera makes these types separable. Since the sensor image is a projection of the ray-space along angular dimensions, the sum of these components creates a low frequency glare for a traditional camera. But by inserting a high frequency occluder, in front of the sensor, these components are converted into a high frequency 2D pattern and can be separated.

The first step of the algorithm for glare reducing is the capture of a 2D high dynamic range photo with a portable light field camera. In the second step the 4D light field gets reconstructed. For each spatial sample of the light field, robust statistical methods are used to eliminate outlier values among its corresponding angular samples. The last step is the reconstruction of the low resolution 2D image by averaging the remaining angular samples for each spatial sample.

6 Photo Collections

Large photo collections and communities are widespread over the internet. These allow to upload, view and share photos and photo albums. New research topics use this communities as photo database for intelligent purpose. One approach is to automatically extract the three dimensional positions and viewing directions of the cameras and find three-dimensional paths through the worlds photos where the user can interactively move along these paths. For example if 3000 photos are made from different perspective of the Statue of liberty, then a software recover the camera poses of all photos and generates controls for browsing through the different views of the statues. Another approach could use large photo collections for completion of a scene. With the intelligent use of a photo database, disturbing obstacles in the photos foreground, can be cut away and seamlessly replaced with the missing part of the background by another photo which is automatically found in the database by the software.

6.1 3D Browsing

Flickr or another large photo communities/databases on the internet have a large amount of photo tourism data from many places of the world. These databases are unstructured and such vast, that the user is not able to get an overview about all photos. For example a google image search on "Notre Dame Cathedral" returns over 15000 photos, capturing the scene from different viewpoints, levels of detail, lighting conditions, season, decades, and so forth.

Snavely et al. [Snavely et al. 2006], [Snavely et al. 2008] developed a 3D photo collections which automatically recover the camera poses (camera positions, viewing directions and field of view) of each photo of these unstructured image collections. In this photo collections the user can browse in 3D, walk through automatically generated orbits or paths. The 3d interface Snavely et. al [Snavely et al. 2006] developed uses image based modelling techniques to recover all viewpoints of each photograph as well as image-based rendering techniques. Smooth transitions between the photographs enable a full 3D navigation for the image collections.

The goal of image-based rendering of this approach, which underlies photo tourism [Snavely et al. 2006], is to create interactive, photo realistic 3D visualizations of real objects and environments. Scene specific controls are generated by analyzing the distribution of the camera viewpoints, the appearance and the scene geometry. The software takes as input a set of photos with a variety of viewpoints, different viewing directions, conditions, cameras and foreground people and outputs an interactive 3D browsing experience with automatic controls, exposed orbits, panoramas, interesting views, optimal trajectories specific to that scene and the distribution of input views.

Scene reconstruction The scene reconstruction algorithm implemented by Snavely et al. [Snavely et al. 2006] uses no additional input data except the image data itself. To reconstruct the camera parameters, firstly feature points of all photos are detected and then they are matched in pairs. Finally an iterative, robust "Structure from Motion" (SfM) procedure is run to recover the camera parameters. Because SfM only estimates the relative position of each camera, and it is necessary to find absolute coordinates (e.g., latitude and longitude), an interactive technique to register the recovered cameras to an overhead map is used. The SIFT keypoint detector with its local keypoint descriptor was used to find and detect keypoints in the input images, which is especially useful for this application because its invariance to image transformations. For each pair of images the approximate nearest neighbors are searched to estimate a fundamental matrix based on RANSAC. In each RANSAC iteration a candidate fundamental matrix is computed using the eight-point algorithm followed by non-linear refinement. Finally, matches which are outliers to the recovered fundamental matrix are removed. If the number of the remaining matches is less than twenty, all of the matches are removed from consideration.

After a set of geometrically consistent matches between each image pair is found, the matches are organized into tracks, where a track is a connected set of matching keypoints across multiple images. Then inconsistent tracks are filtered by removing tracks with less than 2 keypoints. Then for each track the set of camera parameters and the 3D location is recovered. The reprojection error problem, defined as the sum of distances between the projections of each track and its corresponding image features, is solved by algorithms such as Levenberg-Marquardt, which guarante to find local minima.

Figure 16 shows camera viewpoints which are reconstructed from a set of photos from the State Of Liberty taken from Flickr.

The goal of this project is to construct a browsable specific collections of photographs in a 3D spatial context that gives a sense of the geometry of the underlying scene. This approach therefore uses an approximate plane-based view interpolation method and a non-photorealistic rendering of background scene structures.

Automatic path generation The extension of this approach by Snavely et al. [Snavely et al. 2008] enhances the interactive browse experience in 3D. Certain paths of a scene are again generated through a vast number of photos from a community or personal photo collection. The software also reconstructs camera viewpoints from the photo collection and automatically computes orbits, panorama views, canonical views and optimal paths between views. They use image based rendering with a new technique for selecting and warping images for display as the user moves around the scene. This approach maintains a consistent scene appearance. One important aspect of their work is the navigation, which provides control, that make it easy to find interesting aspects of a scene. The key feature of these controls are, that they are generated automatically from photos through analysis of the distribution of recovered camera viewpoints and three dimensional feature distributions using novel path fitting and path planning algorithms.

Figure 16 shows reconstructed camera viewpoints from hundreds of Flickr photos of the Statue Of Liberty and two automatically generated orbits which creates scene-specific controls for image-based rendering and browsing the photo collections.

The components of the extended system are:

- A set of input images and camera viewpoints
- Image reprojection and viewpoint scoring functions
- Navigation controls for a scene
- A rendering engine for displaying input photos
- A user interface for exploring the scene

Reprojection quality The software is able to reproject input images to synthesize new view points and evaluate the expected quality of the reprojection with a reprojection score function. This function should ideally measure the difference between the synthesized

Figure 16: Paths through photo collections. Reconstructed camera viewpoints from hundreds of Flickr photos (top) of the Statue of Liberty and two automatically generated orbits (bottom) which creates scene-specific controls for browsing photo collections in 3D. [Snavely et al. 2008]

view and a real photo of the same scene captured in this view. Because the real photo does not exist, three criteria are used instead for measuring the reprojection quality. At first, the angular deviation is defined as the relative change in viewpoint between the image and the synthesized view. Secondly the projected image should cover as much of the field of view as possible in the synthesized output view. The resolution, which defines the third measure, should be sufficient to avoid blur after projecting. Each of these criteria is scored on a scale from 0 and 1 and defines together the reprojection score. The viewpoint score is defined as the maximum reprojection score for one given synthesized view for all existing input images.

Scene specific controls The software generates scene specific controls because certain types of controls naturally work well for certain types of content. A second reason for scene specific controls is, that different parts of the scene are interesting. For instance, in a virtual art museum, a good set of controls should guide the user from one painting to the next. The software supports three different navigation modes:

- The free-viewpoint navigation allows users to move free around the scene (6-DOF), which gives users the freedom to move wherever they choose, which is limited to the available photos.
- The constrained navigation using scene-specific controls, which allows orbit and panorama controls. Each of such controls is defined by its type, a set of viewpoints and a set of images associated with that control. For an orbit control a set of viewpoints is given on a circular arc with a certain radius looking at a single point.
- The optimized transition from one part of the scene to another guides the user along a automatic generated path to the destination which is associated with a canonical image.

When the scene is already reconstructed, the system automatically analyses the recovered geometry to generate interesting orbits, panoramas and canonical images.

Generate optimized orbits, panoramas and canonical images Optimized orbits have to fulfill several criteria to get detected. They should maximize the quality of the rendered views, span arcs with large angles, have views oriented towards the center of the orbit, which should be a solid object. These objectives of orbit detection involves 1) defining a suitable objective function, 2) enumerating and scoring candidate orbits, and 3) choosing zero or more best-scoring candidates. The objective function is defined as a sum of individual view scores along certain sample positions on the arc. The goal is to maximize this sum, which leads to an optimized orbit. Good orbit axes are computed by finding axes in 3D on which many database images are converged to find the candidate orbits. The objective functions are evaluated along this candidate orbits using the input images from the database. Finally the orbit with the highest score is selected, and all orbits with less than 0.5 score of the selected orbit are removed.

A **panorama** consists of a set of images taken close to a nodal point, which has optimally good views available from a wide range of directions. A panorama scoring function decides how good one image fits to the center of a panorama. The function is evaluated for all images. Candidates with a score beneath a certain threshold are removed.

Canonical images are found by seeking to capture the essence of the scene through a small set of representative images that covers the most popular viewpoints. This method is based on image clustering which matches certain SIFT features.

Rendering When the user moves through the scene, first at all the image with the highest reprojection score is selected to use the best image suited to this viewpoint. If no image has a non-zero score for the current viewpoint, only the point cloud is displayed. Because the image database does not include all viewpoints and directions the rendering engine has to warp the selected image to better match with the current viewpoint. The image gets warped by projecting the original image to a plane which is parallel to the image plane. The system renders the image by projecting it back to the virtual view. Problems occur if the user moves fast through a wide range of views. If for example the user walks through an orbit, the centered object of interest which is associated with the orbit center point may jump around. This unstable alignment is caused through warping and can be avoided through choosing the projection plane so that it intersects the orbit center point and through orienting the rendered view so that the viewing ray which passes the orbit point is always projected in the same pixel position in all views. Reproducing images from other viewpoints is also possible with 3D models extracted from images. A multi-view stereo method for reconstructing geometry from images on flickr exists, but leads to artifacts like projecting foreground objects onto the geometry or to holes in the model. Before the currently selected image is reprojected and rendered onto its proxy plane a background layer, consisting of the reconstructed point cloud, is drawn. The system does alpha fading between instantaneously switches between images.

Moving between paths The system provides controls to move the user automatically along a path from one point in the scene to another by multiple display of the image during traveling between start- and endpoint. For **finding the optimized path** between two points the scene is considered as a transition graph whose vertices are the existing camera samples. Between every pair of images a weighted edge exists which is computed through a transition and a smoothness cost function. Djikstra's algorithm is used to compute the shortest path in the graph. To animate the camera smoothly, a more continous path is produced through smoothing the initial path.

Appearance stabilization The system applies also appearance stabilization methods to compensate the different properties of an unstructured set of photos, for example different lighting conditions(night, day, sunny, cloudy) and different exposures. The visual similarity mode enables the user to select preferred lighting conditions (e.g., only night) and only displays images according to these conditions.

6.2 Scene Completion

Motivation Hays et al. [Hays and Efros 2007] developed a new image completion algorithm which is based on a huge database of photographs gathered from the web. It patches up holes in images by finding similar image regions in a database. The aim is to generate a seamless and semantically valid output image. The algorithm is only driven by data and no labeling of the images is required, see Figure 17.

Scene Completion Strategies There are two different strategies for image completion. In the first strategy the goal is to reconstruct, as accurately as possible, the data that should have been there and to fill in the missing pixel area.

The second strategy of image completion is to hallucinate data that could have been there. The quality of this approach is harder to quantify. The evaluation relies on human visual perception studies. The most successful existing method of this strategy is to extend adjacent textures and contours into the unknown region. The content to fill the unknown parts is taken from the input image, which leads to the problem that mostly the source image does not provide enough data.

The method of Hays et al. [Hays and Efros 2007] focuses on the first strategy which is confronted with three challenges. At first, the high dimensionality of the features from the textures being searched causes a slow texture searching process. Scenes have to be represented by a low dimensional descriptor to accelerate the search of the nearest neighbor images. The second challenge is to find not only a locally matching but also a semantically valid image segment to avoid phenomenons like swimming ducks in city pavement which locally resembles the lake. The third challenge is to fill the missing part with the right color and illumination to avoid a seam between the source image and the new image area. A robust seam finding and blending method, to make the image completions plausible, is needed.

Semantic matching The approach of Hays et al. [Hays and Efros 2007] uses the gist descriptor which is described in the work of Oliva and Torralba [Oliva and Torralba 2006]. The first challenge is handled by acquiring semantic knowledge from the data directly through the gist scene descriptor. The gist scene descriptor is a low-level scene descriptor which is able to group semantically similar scenes(e.g., cities, tall buildings, office, fields, forests, etc.). The big advantage of the usage of the scene descriptor is the low dimensionality which speeds up the search dramatically. Instead of looking for image patches in the whole photo database with more than two million
s of images, the descriptor eliminates 99.99% of the database by finding the nearest neighbor scenes. PCA cannot be used for dimensionality reduction of the images, because the dimensions of the descriptor depending on which regions of a query image are missing. The descriptor is small enough, that a test run with a cluster of 15 machines consisting a image database of about 2.3 million unique images with a total size of 396 gigabytes of JPEG compressed data, which searches semantically valid neighbor images, takes a few minutes.

The gist descriptor aggregates oriented edge responses at multiple scales into very coarse spatial bins. The most effective scene descriptor is build from six oriented edge responses at five scales aggregated to a 4x4 spatial resolution. Color information downsampled to the spatial resolution of the gist is used to augment the scene descriptor. Given an input image the gist descriptor is computed according the missing regions excluded. Therefore a mask is created, which weights each spatial bin in the gist in proportion to how many valid pixels are in that bin. In the next step the distance between the gist of the query image and every gist in the database is computed. Also the color distance is computed, which contributes to the final distance as well.

Figure 18 shows possible semantical matching scenes, by searching for nearest neighbors according to the gist scene descriptor. The next step is to only keep that images, which also match locally. To combine the images seamlessly. Poisson blending is used. To avoid blending artefacts, a graph cut segmentation is performed to find the boundary for the Poisson blending, that has the minimum image gradient magnitude. Minimizing the seam gradient gives the most perceptually convincing compositing result.

Local matching The first step of local context matching is to align the semantical similar scenes to the local image context around the missing region using traditional template matching. The local context is defined to all pixels within an 80 pixel radius from the hole's boundary. This context is compared to the best 200 matching scenes across all valid translations and 3 scales using a pixel-wise error function. The best translations and scales are chosen by minimizing the error function. Additionally to the pixel wise alignment score, also a texture similarity score to measure the coarse compatibility of the proposed fill-in region to the source image within the local context is computed.

The second step of local context matching is to composite each matching scene into the incomplete image at its best placement using a form of graph cut seam finding and standard poisson blending. The seam-finding operation removes valid pixels from the query image. To avoid, that too many pixels are cut, a cost for removing each pixel in the query image is added, which increases with distance from the hole. Because humans are not sensitive to relatively large shifts in color and intensity as long as the shifts are seamless and low frequency, the algorithm tries to minimize the gradient of the image difference along the seam. The seam can be found by minimizing a cost function.

In the final step the poisson blending is applied. Finally a score is assigned to each composite, which is the sum of the scene matching distance, the local context matching distance, the local texture similarity distance and the cost of the graph cut. The 20 best composites results with the lowest scores are presented to the user, where the user can choose one.

This algorithm requires over an hour to process one input image on a single CPU (50 min for scene matching, 20 min for local context matching, and 4 min for compositing). However, by parallelizing the processing across 15 CPUs, it is possible to bring the running time down to about 5 minutes.

After the process is finished the bestmatching image completions are presented to the user. In Figure 19 two examples of image completions, where in each example the original image, the input image with the cut region and three alternative completions are shown. Afterwards the user can select the one he find most compelling.

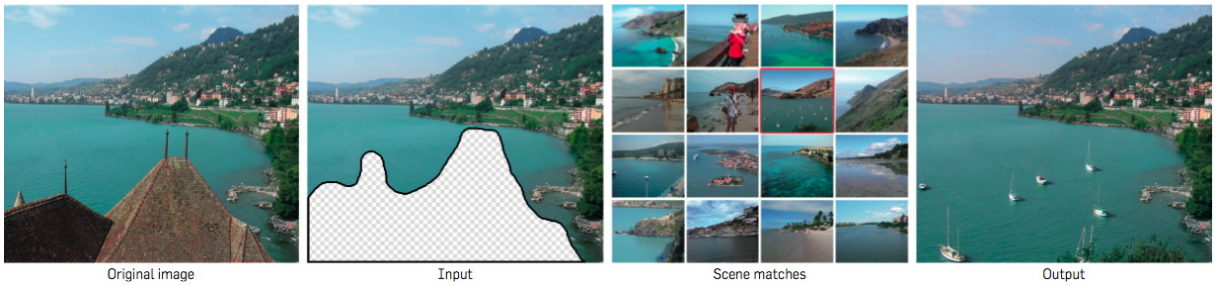

Figure 17: The image completion algorithm tries to fill missing regions in images by finding similar image regions in a large database. In this example the algorithm is used on an image with a cut away obstacle in foreground to replace it with a semantically equivalent image region seamless from another image which is automatically found in the database. [Hays and Efros 2007]

Figure 18: Semantic Matching: The first 164 nearest neighbor scenes for the incomplete image in the center. Most of the scenes are semantically and structurally similar. [Hays and Efros 2007]

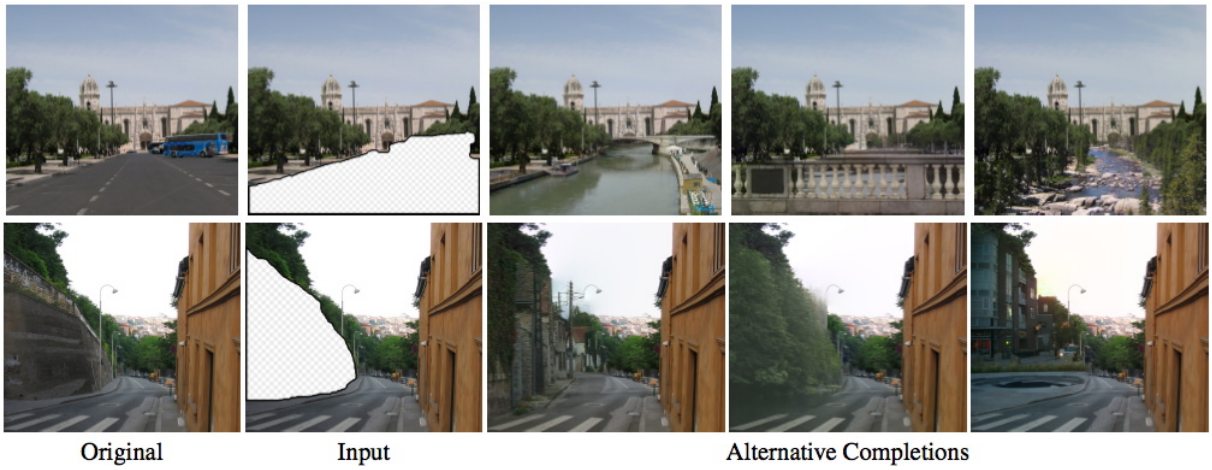

Figure 19: Two examples (top row and bottom row) of compositing with different results. First column: The input image. Second column: The input image with a user defined cut region. Third - fifth column: Three compositing results returned from the system developed by Hays et al. [Hays and Efros 2007], where the user can select the most convincing result.

7 Conclusion

Nobody knows what there will be in tomorrow's digital photography for sure. But we have seen a variety of technologies and recent research topics, which may influence the development of hard- and software in tomorrow's digital photography. We can separate the discussed techniques into two categories:

- Developments for single photos
- Developments for large photo collections

Recent research topics, which were shown in this article, for single photos, have in common that they all have to process several input photos to deliver one output image. This output image has interactive adjustment possibilities after the photo is taken.

Examples are:

- High dynamic range images are computed from several input images with different exposures. To compute one single high dynamic range output image, the tones have to be remapped, such that they can be viewed on current output devices. As we have seen, high dynamic range images are useful for complex lighting situations, where a large range of brightness (from sunlight to shadows) can be captured, to fit better between the visual experience of the captured image and the visual experience of the the reality.

- Today's consumer camera is not able to capture such high resolutions up to gigapixel with a wide field of view. High resolutions, panorama images are therefore also computed from several input photos. Navigation and zooming through the giga pixel panorama image are the described interaction methods.

- Flash matting processes actually two input images, i.e., a flash/non-flash image pair. One interaction method with this approach may be the selection of an new background image, which replaces the original background. This approach needs two input images, while the previous two technologies needs multiple shots due limited camera hardware.

- Light field photography processes the light field, which could also be interpreted as a set of input photos to compute one single output image. The interaction on with possibilities an image captured with a hand-held light field camera shown in this article, are refocussing, sub-aperture adjusting, viewpoint change, glare removal and operations based on the acquired depth field. Current plenoptic camera prototypes suffer from less resolution caused by the microlens array. With the use of larger microlens arrays with smaller microlenses, it is possible to increase the resolution. The programmable aperture approach does not have the loss of spatial resolution, but is restricted to static scenes.

This shows, that the hard- and software of digital cameras has to be improved, to implement these features on tomorrow's camera in a feasible way. As seen above, the possibilities of interactivity compared with today's camera may strongly increase in the future.

Tomorrow's digital photography concerns also the interest of large photo collections. Internet photo collections may be revolutionized with a automatically generated 3D interface. All camera poses can be recovered and set in a three dimensional context, where the user can navigate. Tomorrow's image editing programs might implement scene completion shown in this article. A large photo collection is used, to automatically find a suitable, seamless replacement for a cut-away object of an arbitrary input image. Tomorrow's cameras might be able to establish a connection to the internet which would allow browsing internet photo collections or personal photos in 3D or the complete scenes directly with the user interface on the camera device.

References

CHUANG, Y.-Y., CURLESS, B., SALESIN, D. H., AND SZELISKI, R. 2001. A bayesian approach to digital matting. In *Proceedings of IEEE CVPR 2001*, IEEE Computer Society, vol. 2, 264–271.

DEBEVEC, P., REINHARD, E., WARD, G., AND PATTANAIK, S. 2004. High dynamic range imaging. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Course Notes*, ACM, New York, NY, USA, 14.

FATTAL, R., LISCHINSKI, D., AND WERMAN, M. 2002. Gradient domain high dynamic range compression. In *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, ACM, New York, NY, USA, 249–256.

HAYS, J., AND EFROS, A. A. 2007. Scene completion using millions of photographs. *ACM Transactions on Graphics 26*, 3.

KOPF, J., UYTTENDAELE, M., DEUSSEN, O., AND COHEN, M. F. 2007. Capturing and viewing gigapixel images. *ACM Transactions on Graphics 26*, 3, 93.

LEVOY, M., AND HANRAHAN, P. 1996. Light field rendering. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, ACM, New York, NY, USA, 31–42.

LEVOY, M., 2005. Light field photography and videography. Stanford Computer Graphics Laboratory, http://www-graphics.stanford.edu/talks/.

LEVOY, M. 2006. Light fields and computational imaging. *Computer 39*, 8, 46–55.

LIANG, C.-K., LIN, T.-H., WONG, B.-Y., LIU, C., AND CHEN, H. H. 2008. Programmable aperture photography: multiplexed light field acquisition. In *SIGGRAPH '08: ACM SIGGRAPH 2008 papers*, ACM, New York, NY, USA, 1–10.

NG, R., LEVOY, M., BRÉDIF, M., DUVAL, G., HOROWITZ, M., AND HANRAHAN, P. 2005. Light field photography with a hand-held plenoptic camera. Tech. rep., April.

OLIVA, A., AND TORRALBA, A. 2006. Building the gist of a scene: the role of global image features in recognition. *Progress in brain research 155*, 23–36.

RASKAR, R., AGRAWAL, A., WILSON, C. A., AND VEERARAGHAVAN, A. 2008. Glare aware photography: 4d ray sampling for reducing glare effects of camera lenses. In *SIGGRAPH '08: ACM SIGGRAPH 2008 papers*, ACM, New York, NY, USA, 1–10.

SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. 2006. Photo tourism: exploring photo collections in 3d. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*, ACM, New York, NY, USA, 835–846.

SNAVELY, N., GARG, R., SEITZ, S. M., AND SZELISKI, R. 2008. Finding paths through the world's photos. *ACM Transactions on Graphics 27*, 3, 1–11.

SUN, J., LI, Y., KANG, S. B., AND SHUM, H.-Y. 2006. Flash matting. *ACM Transactions on Graphics 25*, 3, 772–778.

WEIDLICH, A., 2008. Tone Reproduction, rendering lecture unit 8. Vienna University of Technology, http://www.cg.tuwien.ac.at/courses/Rendering/.

WILBURN, B., JOSHI, N., VAISH, V., TALVALA, E.-V., ANTUNEZ, E., BARTH, A., ADAMS, A., HOROWITZ, M., AND LEVOY, M. 2005. High performance imaging using large camera arrays. *ACM Transactions on Graphics 24*, 3, 765–776.

ZYGA, L., 2007. Focus images instantly with Adobe's computational photography. physorg, http://www.physorg.com/news111141405.html.