

Chapter 2

Multi-camera Systems for 3D Video Production

2.1 Introduction

As discussed in the previous chapter, 3D video records full 3D shape, motion, and surface texture of an object in motion rather than a pair of stereo video or 2.5D range data. To produce such data, the entire 3D object surface should be captured simultaneously. The practical method for this is to employ a group of video cameras,¹ place them to surround an object in motion, and reconstruct its 3D shape, motion, and surface texture from a group of multi-view video data recording partial 2D or 2.5D object views. While several advanced 3D video capture systems [5] are being developed introducing Time-Of-Flight cameras [25] and/or active-stereo cameras with structured lights to capture 2.5D range video data in addition to ordinary cameras, we do not consider such 2.5D cameras in this book and present 3D video production methods by reconstructing 3D object shape, motion, and surface texture from multi-view 2D video data.

General limitations of current 3D video production technologies are:

- In principle, multiple objects in motion can be captured at the same time. In practice, however, since their mutual occlusions degrade the quality of 3D video data, most of 3D video data are produced for a single object. Thus, in what follows we assume a 3D video stream of one object in motion is produced, except when we explicitly refer to multiple objects.
- Since the problem of reconstructing 3D object shape, motion, and surface texture in natural environments is very difficult due to dynamically changing background objects and lighting environments, most of 3D video data are produced from multi-view video data captured in well-designed studios.

As will be discussed in the next part, even though we assume a single object in motion in a well-designed studio, there remain many technical problems to be solved for producing high fidelity 3D video.

¹In what follows, we simply refer video cameras as cameras.

Table 2.1 Camera parameters and their effects

Parameter	Effect
Iris	The smaller the size, the deeper the depth of field, but the darker the image becomes
Gain	The smaller the gain, the less noisier, but the darker the image becomes
Shutter	The faster the shutter, the less motion blurred, but the darker the image becomes
Zoom	The smaller the zooming factor, the less deeper depth of field and the wider the field of view, but the smaller the image resolution becomes

This chapter presents and discusses requirements, design factors, and implementation methods of a multi-view camera studio for 3D video production (3D video studio, for short).

The basic policy we employed is to implement 3D video studios with off-the-shelf devices rather than develop specialized ones for 3D video production. This is not only to develop cost effective systems for casual usages but also to investigate essential problems in 3D video production. Thus, all devices introduced in this and the next chapters can be easily prepared to start research and development of 3D video.

2.1.1 Single-Camera Requirements

Firstly, the requirements for 3D video studios can be classified into two categories: single-camera requirements and multi-camera requirements. The former include the following.

1. A camera should be kept well focused on the object during its motion.
2. Captured video data should not contain any motion blur even if the object motion is fast.
3. The dynamic range of a camera should be adjusted to lighting environments in a studio to capture color data accurately.
4. The resolution of a camera should be high enough to capture detailed object surface textures.
5. The field of view of a camera should be wide enough to capture an object in motion.

To satisfy these requirements, the camera parameters should be adjusted: focus, iris (aperture size), gain, color balance, shutter speed (exposure time), zoom (focal length, or field of view), and position and orientation (pan and tilt). Table 2.1 summarizes effects by some of these parameters, which show mutual dependencies and hence trade-offs among them. For example, while closing the iris and shortening the exposure time as much as possible are useful to satisfy the requirements 1 and 2 above, very powerful lightings are required to satisfy the requirement 3. Moreover, the requirements 4 and 5 are in a trade-off relation, whose practical solution with

active cameras will be given in Chap. 3. Thus, we need to find an acceptable set of the parameters by considering trade-offs between them.

2.1.2 Multi-camera Requirements

While the single-camera requirements are well known and various types of know-how have been developed in photography and cinema production, multi-camera requirements are rather unique in computer vision and some modern cinematography with multiple camera systems. They include:

1. Accurate 3D positions and viewing directions of multiple cameras should be known to integrate captured multi-view video data geometrically.
2. Multiple cameras should be accurately synchronized to integrate captured multi-view video data temporally.
3. Accurate brightness and chromatic characteristics of multiple cameras should be known to integrate captured multi-view video data chromatically.
4. All object surface areas should be observed by at least two cameras to reconstruct their 3D shapes by stereo-based methods; while visual cues in a single image such as shading can be used to reconstruct 3D object shape, absolute 3D depth cannot be computed and, moreover, many assumptions which are not always valid in the real world are required.

Requirements 1, 2, and 3 imply that cameras should be well calibrated geometrically and photometrically as well as synchronized. While these requirements can be satisfied in some accuracy with modern camera calibration methods, the last requirement is rather hard to satisfy. Especially for objects with loose clothes such as MAIKO and objects playing complex actions such as Yoga, it is not possible to satisfy this requirement. As will be discussed in detail in the next part, moreover, the multi-view surface observability plays a crucial role in the 3D shape and motion reconstruction (Chap. 4) and the texture generation (Chap. 5) for 3D video production. Consequently, the layout design of cameras should be done carefully to allow as many object surface areas as possible to be observed.

As the first step toward 3D video production, this chapter establishes technical understandings about how we can find a feasible set of camera parameters which satisfies the above-mentioned requirements in practice. Section 2.2 first discusses design factors of 3D video studios and introduces three 3D video studios we developed. Then, Sect. 2.3 presents geometric and photometric camera calibration methods. The introduction and calibration of active cameras for tracking and capturing multi-view video of an object moving in a wide spread area will be presented in Chap. 3. The calibration or estimation of lighting environments will be presented in Chap. 6, since it has much to do with the texture generation in Chap. 5. Section 2.4 evaluates the performance of the three 3D video studios we developed, where the accuracy of the geometric camera calibration is quantitatively evaluated. Section 2.5 concludes this chapter with discussions and future works.

2.2 Studio Design

Here we discuss technical problems to design a 3D video studio with *static* cameras. While, as will be discussed later, such system constrains the object movable space to satisfy the requirements described above, most of 3D video studios developed so far used static cameras to produce high quality 3D video. The introduction of *active* cameras, which cooperatively track an object and capture multi-view high-resolution video, is a promising method to expand the object movable space. Since such an active multi-view video capture system should satisfy an additional requirement for dynamic camera control synchronized with the object motion, we confine ourselves to static cameras in this chapter and reserve discussions on active cameras for Chap. 3.

2.2.1 Camera Arrangement

One of the most important design factors of a 3D video studio is how to determine (1) the number of cameras to be installed and (2) their spatial arrangement to achieve high 3D shape reconstruction accuracy. If we do not have any specific knowledge about the object shape or motion, or if we want to capture a variety of objects in the same studio, one reasonable solution is to employ a circular ring camera arrangement, where a group of cameras placed evenly along the ring observe the object performing actions at the ring center. We may call it a *converging multi-camera arrangement*. Figure 2.1 illustrates three typical multi-camera arrangements: *diverging multi-camera arrangement* for omni-directional image capture and *parallel multi-camera arrangement* for multi-baseline stereo [23, 27] and light-field modeling [15].

Then the next design factors to be specified concern the placement of the camera ring and the number of cameras installed on the ring. In [28], we pointed out:

- The best observability of the object surface with a single ring camera arrangement is achieved by locating the ring at the mid-height of the target object.
- Shape-from-silhouette methods for 3D shape reconstruction (Sect. 4.2.2.2) require at least nine cameras (40° spacing on the ring), and the reconstruction accuracy can be improved well by increasing the number of cameras up to 16 (23°). Even with larger number of cameras, the accuracy improvement is limited, since shape-from-silhouette methods can only reconstruct an approximated 3D shape of the object by definition (cf. “visual hull” in Sect. 4.2.2.2).
- Shape-from-stereo methods require at least 14 cameras (25°) for an optimal balance between matching accuracy and depth ambiguity; the wider the baseline between a pair of stereo cameras becomes (i.e. wide-baseline stereo), the better accuracy of the depth measurement is achieved, while the harder the stereo matching becomes.

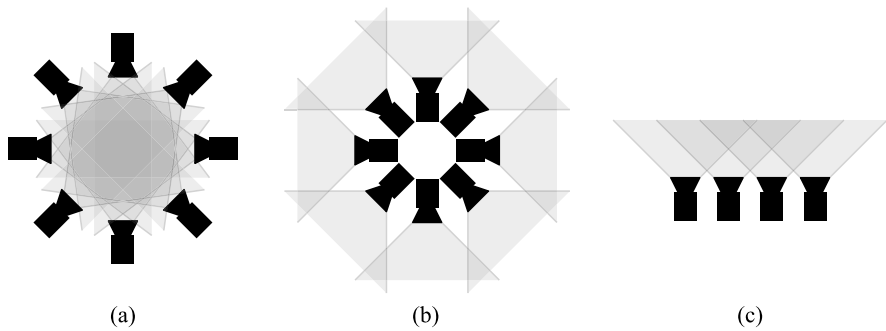


Fig. 2.1 Multi-view camera arrangements: (a) converging, (b) diverging, and (c) parallel arrangements

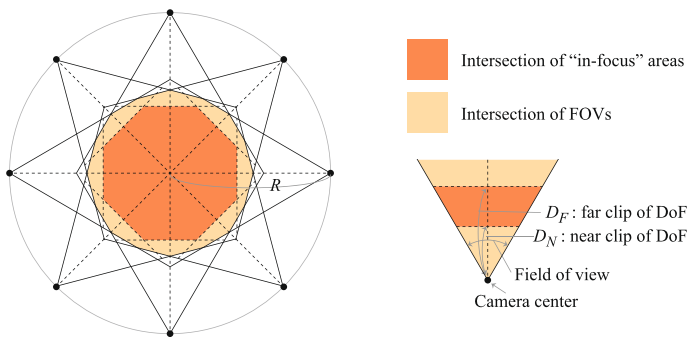


Fig. 2.2 Capturable area. Each camera can capture an object located in its field of view and within the depth-of-field (DoF) without blur. The capturable area of a multi-view camera system is given by the intersection of such “in-focus” areas

Hence, we conclude here that we need at least nine to 16 cameras for a 3D video studio with a single ring camera arrangement. As will be shown later, in Sect. 2.2.7, practical 3D video studios are usually equipped with ceiling cameras in addition to a camera ring(s) to increase the observability of top areas of an object.

The camera arrangement constrains the object movable space to guarantee the multi-view observability of the object surface. In general, the 3D observable space of a camera can be represented as a quadrilateral pyramid formed by its projection center and bounded image plane. Thus, intuitively, with a converging multi-camera arrangement, the object movable space is confined within intersections of multiple quadrilateral pyramids (Fig. 2.2). That is, to guarantee the surface observation by at least two cameras, the object can only move in spaces where at least two quadrilaterals intersect. Similar space limitations are introduced also by focusing and zooming. They will be described later in this section.

It should be noted that with enough number of cameras, all of them do not need to capture entire object images. That is, as long as all object surface areas can be observed by multiple cameras, some of cameras can capture the object partially by

Table 2.2 Categorization of video cameras

	Media production	Machine vision	Consumer
Cost	High	Middle to low	Middle to low
Quality	High	Middle to low	Middle to low
Data transmission	HD-SDI	IEEE1394b, 1000Base-T, CameraLink, USB3.0	USB, IEEE1394a
Synchronization	GenLock + Timecode	Trigger Signal	N/A
Lens	PL- or PV-mount	C-, CS-, or F-mount	Unchangeable

zooming up to increase image resolution. In fact, one of our studios (Studio B in Table 2.3) employed this strategy to increase image resolution.

2.2.2 Camera

A large variety of commercial cameras are available in the market. They can be categorized by their application domains (Table 2.2). The first group is for professional media productions designed to achieve high-end quality: high-resolution and high color-depth. The second group is for industrial and machine vision. They are originally designed for factory automation, robot, etc., and relatively low-cost. The last group is for consumer use. They are widely available in the market, but not fully designed to interoperate with other cameras or controllers. Since 3D video studios require the synchronization of multiple cameras, consumer cameras cannot be used.

The important difference between media production and machine vision cameras is twofold. The first is in their image qualities. Since media production cameras typically utilize 3CCD system, they offer full 8-bit depth for each color channel. On the other hand, most of machine vision cameras utilize 1CCD system with Bayer color filter [2], and their effective color-depth is reduced into 1/3.

The second difference is in their synchronization mechanisms. While both media production and machine vision cameras accept a signal to control video capture timing, there is an important difference in the temporal structures of timing signals allowed. In the GenLock (generator lock) system for media production cameras, the signals should come regularly with a standardized interval such as 24 Hz, 29.97 Hz, etc. On the other hand, trigger systems for machine vision cameras allow signals to arrive at arbitrary timings. When selecting cameras, these two different synchronization mechanisms should be taken into account, especially when both types are employed into a 3D video studio. Note that some of machine vision cameras have yet another synchronization mechanism called “bus-sync”. It makes all cameras on the same bus synchronized automatically without providing additional signals.

Other practical factors when selecting cameras are the allowable cable length and the data transmission rate between a camera and its data receiver. HD-SDI (formally SMPTE-292M) connection for media production cameras and 1000Base-T

machine vision cameras (known as “GigE Vision” cameras standardized by AIA) allow 100 m cable length. On the other hand, IEEE1394b (or FireWire800), CameraLink, and USB3.0 connections for machine vision cameras allow only 3 m to 10 m without active repeaters. Note that some non-standard long cables are available in the market. Thus, the camera selection for real time multi-view video capture should be done taking into account the physical size of the 3D video studio, the bandwidth of video data transfer, and the processing speed of computers and storage devices.

2.2.3 Lens

While not discussed usually, the lens selection is very important to guarantee high quality multi-view image capture, because a lens specifies the field of view, the amount of incoming light, and the depth of field.

The *field of view* can be computed from the physical imager size and the effective focal length of the lens. Suppose the imager size is W mm \times H mm and the effective focal length is f mm. Then the horizontal and vertical field of view angles are simply given by

$$\begin{aligned} \text{FOV}_H &= 2 \tan^{-1} \left(\frac{W}{2f} \right), \\ \text{FOV}_V &= 2 \tan^{-1} \left(\frac{H}{2f} \right). \end{aligned} \tag{2.1}$$

Imager sizes are often described by their “format”, such as “1/1.8 inch sensor”. For some historical reasons in optics, this number is equal to the diagonal size of the imager divided by 16; that is, the diagonal length of “1/1.8 inch sensor” is $1/1.8 \times 16 = 8.89$ mm.

The amount of light recorded by an imager through a lens is denoted by *F-number* (or F-ratio, F-stop). The *F-number* is a dimensionless value given by the focal length divided by the effective aperture diameter of the lens. The larger the *F-number*, the smaller the lens opening is, and the lesser light comes in. Therefore it is better to use a lens with smaller *F-number* to capture brighter images of scenes under limited lighting environments.

F-number also specifies the *depth of field* of a lens, which defines the depth range in which images can be captured without blur. A small *F-number* means a small depth of field. Since the physical pixel size is the finest resolvable point size in an image, blurring within this size does not introduce any effects in a captured image. This size is known as the *circle of confusion*, the maximum tolerable size of blurring. When a lens is focused at infinity, the farthest distance D_H beyond which all object images are not blurred can be computed from the circle of confusion diameter c as follows:

$$D_H \approx \frac{f^2}{F c}, \tag{2.2}$$

where f denotes the focal length and F the F -number. This distance D_H is called *hyperfocal distance*. If the lens is focused at $d_f < D_H$ distance from the optic center, then the nearest and farthest distance between which all object images are not blurred are given as

$$D_N \approx \frac{D_H d_f}{D_H + d_f}, \quad (2.3)$$

$$D_F \approx \frac{D_H d_f}{D_H - d_f}, \quad (2.4)$$

respectively. Hence the depth-of-field is

$$\text{DOF} = D_F - D_N = \frac{2D_H d_f^2}{D_H^2 - d_f^2}. \quad (2.5)$$

For example, let the physical pixel size be $4.4 \mu\text{m} \times 4.4 \mu\text{m}$, the focal length 6 mm, the F -number 1.4, and the focus distance 2.5 m. Then $D_H \approx (6 \text{ mm})^2 / (4.4 \mu\text{m} \times 1.4) = 5.84 \text{ m}$, $D_N \approx (5.84 \times 2.5) / (5.84 + 2.5) = 1.75 \text{ m}$ and $D_F \approx (5.84 \times 2.5) / (5.84 - 2.5) = 4.37 \text{ m}$. This means that when cameras are placed on a ring of 3 m radius, an object located within 1.25 m from the ring center can be captured in good focus without blurs. However, if it moves more than $1.25 \text{ m} = 3 \text{ m} - 1.75 \text{ m}$ to a camera, then the image captured by that camera will not be well focused. That is, for a 3D video studio with a ring camera arrangement of radius R , the capturable area in terms of the depth of field can be approximated by the intersection of concentric circles of diameter $2R - 2D_N$ and $2R - 2D_H$ as illustrated by Fig. 2.2, which further constrains the movable space of an object.

2.2.4 Shutter

The shutter speed controls the amount of motion blur as well as incoming light. By shortening the shutter, we can suppress the motion blur while reducing the amount of incoming light. Similarly to the discussion on the depth of field, if the object motion appears smaller than the pixel size, then the image does not include any effects of motion blur.

There are two different types of shutter: global and rolling shutters. With the *global shutter*, all pixels in the imager start and end exposure simultaneously. In contrast, the *rolling shutter* makes each pixel line start exposure one by one while a captured image can be transmitted frame-wise. This introduces illusionary deformations into dynamic object images, and makes 3D video production unnecessarily harder. Therefore we suggest global shutter cameras, most of which have CCD sensors.

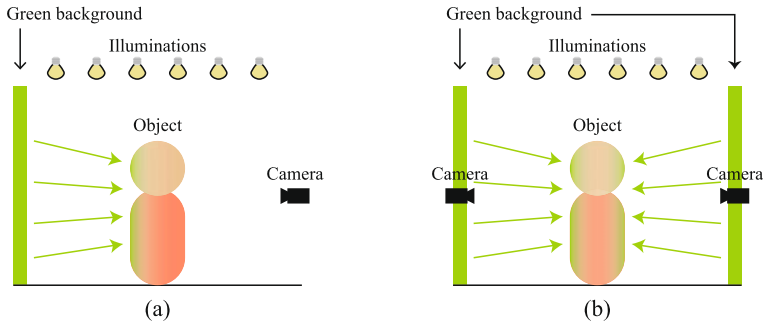


Fig. 2.3 Single-view and multi-view chroma-keying. (a) In single-view chroma-keying, colored reflections from the background to the object surface are occluded from the camera. On the other hand, in multi-view environment (b), colored reflections are observed from multiple cameras

2.2.5 Lighting

In a 3D video studio, the camera arrangement constrains the arrangement of light sources as well as the object movable space. In general, cameras should not observe light sources directly, because strong direct lights damage captured images. While ordinary single-camera systems can avoid this problem by locating light sources next to the camera, such light source arrangement cannot be used for multi-view ring camera systems; a light source placed near by a camera is captured by other cameras. Thus, one reasonable solution is to locate light sources on the ceiling and set viewing directions of cameras so that captured images do not include the ceiling (Fig. 2.3). To facilitate such light source arrangement, 3D video studios should have enough heights as ordinary TV studios and theaters.

As discussed before, to enhance multi-view image capture capabilities of a 3D video studio, the amount of incoming light to an image sensor is reduced (1) with a smaller iris to make the depth-of-field wider and the capturable area wider, and (2) with a shorter shutter speed to avoid motion blur. To compensate for these darkening effects, we should increase lighting or the sensor gain, which usually reduces the SN ratio of captured images.

Typical lighting systems consist of halogen lamps, fluorescent tubes, LEDs, etc. While they have different characteristics on their initial cost, energy efficiency, life time, color, and so on, an important point for the 3D video studio design is whether it does flicker or not. In particular fluorescent tubes without inverters blink at 100 or 120 Hz (double of AC input), and make the global illumination level drift periodically. This should be avoided in a 3D video studio.

Besides these continuous lighting devices, we can use lighting devices which flash synchronously to camera exposures. For example, we can use projectors as programmable lights, or strobe lights to “freeze” object images in quick motion [31]. To make full use of such dynamic lighting, well-designed synchronization controls should be developed to coordinate video capture and lighting.

Another augmentation of lighting is the introduction of structured lights [1, 26] to realize active-stereo analysis. Since high beams of structured lights may disturb human actions to be captured, infra-red structured lighting systems are used. In fact, advanced 3D video systems being developed [5] employ such active sensing devices in addition to ordinary cameras.

While studio and theater lighting designs have been well studied and effective lightings are very important design factors to produce attractive visual contents, this book does not cover them except for Chap. 6, which presents a method of estimating 3D shapes, positions, and radiant intensities of distributed dynamic light sources.

2.2.6 Background

As will be discussed in Chap. 4, multi-view object silhouettes are very useful for the 3D object shape reconstruction. In particular, the accurate silhouette contour extraction is very crucial, since it directly defines the accuracy of the visual hull geometry (Sect. 4.2.2.2). In fact, the visual hull is often used as the initial estimation of the 3D object surface in practical algorithms (Sect. 4.4).

One straightforward solution for the silhouette extraction is to employ background subtraction or chroma-keying techniques. In the former, an object silhouette is given as the difference between a captured object image and the background image taken beforehand without any object. In the latter, on the other hand, the background with a known uniform color is prepared and an object silhouette is extracted as image regions having colors different from the background color. Both techniques are well studied and produce images in media production quality for studio setup.

However, it should be noted that the chroma-keying for multi-view camera studio introduces non-negligible color bias into captured images (Fig. 2.3). That is, blue or green lights reflected from the background illuminate the object. In single-view chroma-keying, widely used for cinema and broadcast media production, this is known as “blue (or green) spill”. It appears typically only around the occluding boundary, because most of the reflected lights are occluded by the object. In 3D video studios, on the other hand, all surface areas are lit by colored reflections from the background. To avoid this color bias, we can use the gray background as used in Studios A and B in Fig. 2.4, or estimate lighting environments in a 3D video studio by such methods as presented in Chap. 6 and neutralize the illumination bias. The latter approach is left for future studies.

While we do not discuss object silhouette extraction methods in this book, even with the state-of-the-art computer vision technologies, it is still not possible to achieve the perfect accuracy. Especially, when an object wears very colorful clothes like MAIKO with FURISODE, the chroma-keying does not work well and, moreover, wrinkles of her loose FURISODE are covered with soft shadows, and decorations in gold thread generate highlights. To cope with such complicated situations, ordinary 2D image processing methods alone are not enough and hence advanced

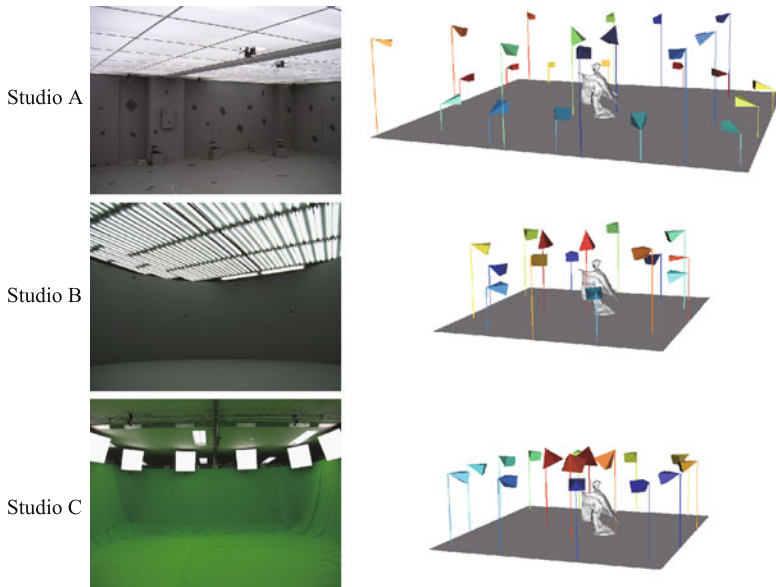


Fig. 2.4 Three 3D video studios developed at Kyoto University. The *left column* shows their interior scenes and the *right* the camera arrangements, respectively. The *colored quadrilateral pyramids* in the camera arrangements illustrate the projection centers and fields of view of the cameras

methods which integrate both the multi-view 2D silhouette extraction and the 3D shape reconstruction should be developed [8, 9, 13, 29, 32].

In summary, the problem of 3D video studio design can be regarded as the optimization of the object surface observability by a group of cameras, i.e. the surface coverage by multi-view images of well-focused, high spatial resolution, and high fidelity color. Since an object freely moves and performs complex actions, it is not possible to compute the optimal design analytically. Chapter 3 derives algebraic constraints in designing a 3D video studio with active cameras and analyzes their mutual dependencies to obtain a feasible solution. Finally, it should be noted that the 3D video studio design should be done based on real world physics, while the camera calibration discussed below is conducted based on a simplified algebraic model.

2.2.7 Studio Implementations

Figure 2.4 and Table 2.3 show three 3D video studios and their specifications we developed so far, respectively. They were designed for different objectives.

Studio A was designed to develop a 3D video studio with multi-view active cameras, which track and capture an object moving in a wide spread area. Its computational algorithm and technical details will be presented in Chap. 3.

Table 2.3 Specifications of three 3D video studios developed at Kyoto University

	Studio A	Studio B	Studio C
Feature	Wide area	Accurate shape and color	Transportable
Shape	Square	Dodecagon	Rounded square
Size	10 m × 10 m 2.4 m height	6 m diameter 2.4 m height	6 m diameter 2.5 m height
Camera	high and low double rings	high and low double rings	single ring
Arrangement	with ceiling cameras	with ceiling cameras	with ceiling cameras
Camera	Sony DFW-VL500 × 25	Sony XCD-X710CR × 15	Pointgrey GRAS-20S4C × 16
Imager	1/3 inch 1CCD	1/3 inch 1CCD	1/1.8 inch 1CCD
Image format	VGA/RAW	XGA/RAW	UXGA/RAW
Lens	Integral 5.5 mm to 64 mm	C-mount 6 mm & 3.5 mm	C-mount 6 mm & 3.5 mm
Pan/tilt/zoom	Active (with pan/tilt unit)	Static	Static
Frame rate	12.5fps	25fps	25fps
Capture PC	25	15	2
Connection	IEEE 1394a 20 m cable	IEEE 1394a 20 m cable	IEEE 1394b 10 m cable
Datarate	3.66 MB/s	18.75 MB/s	45.78 MB/s (366 MB/s per PC)
Background	Gray plywood	Gray plywood	Green screen
Lighting	Overhead inverter fluorescent lights		

Studio B was designed to produce 3D video with accurate object surface geometry and texture for digital archiving of Japanese traditional dances. Most of multi-view video data used in this book were captured in this studio. Its gray static background eliminates the color bias discussed before and allows high fidelity colored surface texture generation, which is an important requirement for digital archiving, especially for colorful Japanese clothes, KIMONO. Note, however, that chroma-keying with gray background often introduces errors in object silhouettes: soft shadows at small wrinkles on object clothes are captured as gray regions. To remove such errors, image segmentation and/or 3D shape reconstruction methods should employ the constraints on the connectivity of silhouette regions and the inter-viewpoint silhouette consistency [22].

Studio C was designed as a transportable 3D video studio to realize on-site 3D video capture. To minimize the studio equipments, it employs only two PCs to receive 16 UXGA video streams, and the green screen background for easier silhouette extraction.

2.3 Camera Calibration

Following the 3D video studio design, its geometric and photometric calibrations should be done for obtaining multi-view video data usable for 3D video production.

2.3.1 Geometric Calibration

2.3.1.1 Camera Model

The geometric camera calibration is the process that estimates parameters of the geometric transformation conducted by a camera, which projects a 3D point onto the 2D image plane of the camera. Figure 2.5 illustrates the camera model used in this book. Note that this *pinhole camera model* simplifies geometric transformations conducted by a physical camera and hence cannot represent important physical characteristics required to design a 3D video studio such as the depth of field. While closely related, therefore, the 3D video studio design and the camera calibration should be considered as separate processes.

As shown in Fig. 2.5, the position of a 3D point in the scene is described by vector ${}^Wp = (x, y, z)^\top$ in the world coordinate system W . Wp is transformed to the camera coordinate system C by

$${}^Cp = R {}^Wp + T = (R | T) \begin{pmatrix} {}^Wp \\ 1 \end{pmatrix}, \quad (2.6)$$

where R and T are the rotation matrix and the translation vector which describes the position and posture of the camera in the world coordinate system. Then the point Cp in the camera coordinate system is transformed to $(u, v)^\top$, the ideal position in the image coordinate system without considering the lens distortion:

$$\lambda \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = A {}^Cp = \begin{pmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{pmatrix} {}^Cp = \begin{pmatrix} k_u & s & u_0 \\ 0 & k_v & v_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix} {}^Cp, \quad (2.7)$$

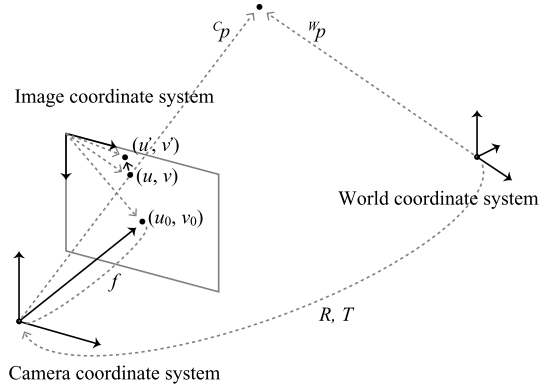
where λ is a scale parameter which normalizes the third component of the left-hand side vector to 1. By definition λ is equal to the z -value (depth) of Cp . f denotes the effective focal length of the camera in pixel. k_u and k_v denote the aspect ratio of the pixel, s denotes the skew parameter, and (u_0, v_0) the intersection point of the optic axis with the image screen represented by the image coordinate system.

Given $(u, v)^\top$, its observed position $(u', v')^\top$, which is transformed with lens distortions, is modeled as a mapping in the normalized camera coordinates:

$$\begin{pmatrix} N_x' \\ N_y' \end{pmatrix} = (1 + k_1 r^2 + k_2 r^4) \begin{pmatrix} N_x \\ N_y \end{pmatrix}, \quad (2.8)$$

Fig. 2.5 Camera model.

A 3D point is first projected onto the ideal position (u, v) in the 2D image plane, and then shifted to the observed position (u', v') by lens distortions



where $r^2 = N_x^2 + N_y^2$. k_1 and k_2 are the radial distortion parameters. The normalized coordinate system is given by

$$\lambda \begin{pmatrix} N_x \\ N_y \\ 1 \end{pmatrix} = C_p. \quad (2.9)$$

In other words, the matrix A in Eq. (2.7) of the normalized camera is the identity matrix. Finally, $(u', v')^T$ is given as

$$\begin{pmatrix} u' \\ v' \\ 1 \end{pmatrix} = A \begin{pmatrix} N_x' \\ N_y' \\ 1 \end{pmatrix}. \quad (2.10)$$

In this camera model, R and T are called *extrinsic parameters*. A is called the *intrinsic parameter* since it is independent of the camera position and posture. k_1 and k_2 are also independent of the extrinsic parameters, but are called *lens distortion parameters* in particular.

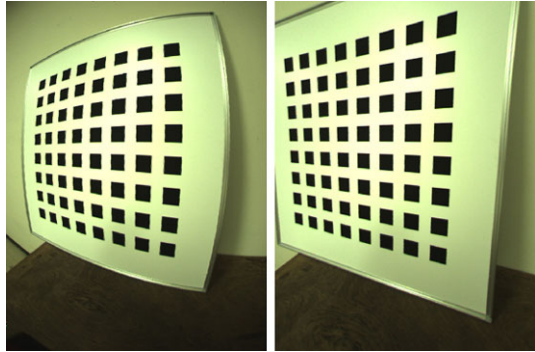
The geometric calibration is a process which estimates these extrinsic, intrinsic, and lens distortion parameters by observing some reference objects in the scene.

2.3.1.2 Computational Methods for Static Camera Calibration

In general,

- The camera calibration should be done by placing reference objects around the 3D local area where an object to be captured in 3D video performs actions. This is because the accuracy of the camera calibration is guaranteed only around the reference objects.
- The camera calibration should employ a non-linear optimization like the bundle adjustment as the final step to minimize a geometrically meaningful error metric such as the reprojection error.

Fig. 2.6 Planar pattern for the camera calibration. *Left*: observed image. *Right*: rectified image using estimated intrinsic and lens distortion parameters



This section introduces a practical four-step calibration procedure while any calibration procedures can be used as long as the above-mentioned points are satisfied:

- Step 1.** Intrinsic and lens distortion parameters estimation by Zhang [33].
- Step 2.** Extrinsic parameter calibration by 8-point algorithm [10].
- Step 3.** Non-linear optimization (bundle adjustment).
- Step 4.** Global scale and position adjustment.

2.3.1.2.1 Intrinsic and Lens Distortion Parameters Estimation

The most standard camera calibration method is a planar pattern-based method proposed by Zhang [3, 4, 33]. Given a set of planar reference 3D points whose positions on the plane are known, it estimates the camera position and posture with respect to the reference, and the intrinsic and lens distortion parameters. Figure 2.6 shows the planar pattern used for the calibration. In this method, the planar pattern defines the world coordinate system.

This method, however, cannot be used in the calibration of the multi-camera system in the 3D video studio.

- With the ring camera arrangement, the placement of the planar pattern is very limited to guarantee the simultaneous observation by all cameras. While a possible placement to satisfy the simultaneous multi-view observation is to place it on the floor, the accuracy of the pattern detection in observed images is degraded because the cameras can observe the plane at very shallow angles.
- The placement limitation can also degrade the overall calibration accuracy; the reference plane should be placed in the object action space to guarantee the calibration accuracy.

Note that a transparent planar pattern would solve these problems, while its specular surface reflections would introduce another placement limitation from lighting environments. Thus, we use Zhang's method only for the intrinsic and lens distortion

parameter estimation, which can be done for each camera independently, and employ a multi-view extrinsic parameter estimation method at the second step.

With Zhang's method, the intrinsic parameters represented by A in Eq. (2.7) and the lens distortion parameters k_1 and k_2 in Eq. (2.8) are estimated. Figure 2.6 compares a captured image of the reference pattern and its rectified image with the estimated parameters.

2.3.1.2.2 Extrinsic Parameter Estimation

Given the intrinsic and lens distortion parameters for each camera, we can compute the relative positions of multiple cameras by linear 8-point [10], non-linear 5-point [20], or trifocal-tensor-based algorithms [6] from 2D-to-2D point correspondences (Fig. 2.7).

To implement a practical extrinsic parameter estimation method, we have to develop methods to (1) obtain accurate 2D-to-2D point correspondences, and (2) calibrate multiple cameras from the 2D-to-2D point correspondences.

For (1), we can make full use of the synchronized multi-view image capture. That is, move a uniquely identifiable reference object(s) scanning the possible object action space. Then, regard reference object positions in simultaneously captured multi-view images as corresponding points. To make this method work well, feature point(s) on the reference object should be designed as view-invariant: for example, 2D chess corners or a center of a 3D sphere (Fig. 2.8).

A simple solution for (2) above is to use the 8-point algorithm for estimating the relative position and posture of each camera pairs. Since the 8-point algorithm estimates only the pair-wise relative position up to a scale factor, we should determine the relative positions of all cameras by the following process. Let us consider three cameras A , B , and C as the minimal setup for multi-camera calibration.

1. Suppose we use camera A as the reference, i.e., we are going to describe positions and postures of B and C in the camera A coordinate system.
2. Estimate the relative position and posture for each pair $A \leftrightarrow B$, $B \leftrightarrow C$, and $C \leftrightarrow A$. Note here that we have unknown scale factors for each pair of cameras: λ_{AB} , λ_{BC} , and λ_{CA} (Eq. (2.7)). Let the relative posture and position of Y w.r.t. X be ${}^X R_Y$ and ${}^X T_Y$ which transforms a point ${}^Y p$ in the camera Y coordinate system to the X coordinate system by ${}^X p = {}^X R_Y {}^Y p + \lambda_{XY} {}^X T_Y$. Here we can assume $|{}^X T_Y| = 1$ without loss of generality.
3. Let $A0$ denote the origin of the camera A coordinate system.
4. The origin of the camera B coordinate system is represented by ${}^A R_B {}^B 0 + \lambda_{AB} {}^A T_B = \lambda_{AB} {}^A T_B$ in the camera A coordinate system.
5. Similarly, the origin of the camera C coordinate system is represented by $\lambda_{AC} {}^A T_C$ in the camera A coordinate system.
6. On the other hand, the origin of the camera C coordinate system is represented by $\lambda_{BC} {}^B T_C$ in the camera B coordinate system, which is represented by $\lambda_{BC} {}^A R_B {}^B T_C$ in the camera A coordinate system. Then, the origin of the camera C coordinate system is represented by $\lambda_{BC} {}^A R_B {}^B T_C + \lambda_{AB} {}^A T_B$.

Fig. 2.7 Extrinsic parameter estimation. With several known 2D-to-2D point correspondences in a pair of observed images (p_1 to p'_1, \dots, p_n to p'_n), the relative 3D position and posture of two cameras (R and T) can be estimated up to scale

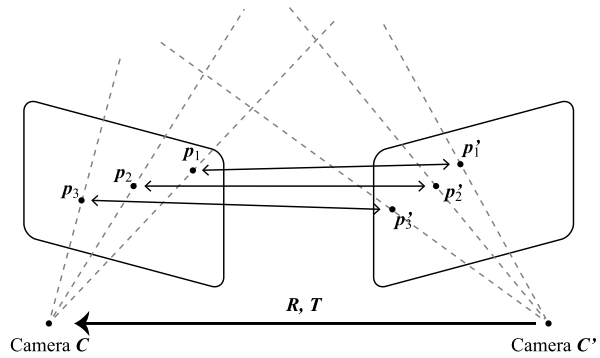
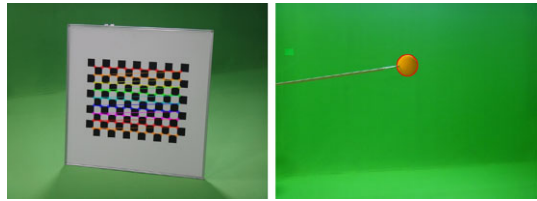


Fig. 2.8 2D-to-2D point correspondences by using chess corners (*left*, by [4]) and sphere centers (*right*)



7. By equating the above two representations of the origin of the camera C coordinate system, we can obtain the constraint for three scale factors. That is, the three coordinate systems of cameras A , B , and C are integrated into the common coordinate system with one scale factor.

By iteratively applying this method for the other cameras one by one, we can describe all the camera positions and postures in the camera A coordinate system with a scale factor.

Notice that this process obviously accumulates calibration errors through the iteration. However, this is not a serious problem since the following non-linear optimization will reduce these errors.

From a practical point of view, we can use this extrinsic parameter calibration to verify whether or not the calibration processes and the multi-camera system are working correctly. That is, if the 8-point algorithm fails to estimate the pair-wise positions and postures, that is, if calibration errors such as sum of the reprojection errors (described in the next section) are not acceptable, then examine if

1. The corresponding point estimation may have introduced errors due to false-positive and/or true-negative detections, or
2. The multi-camera synchronization may not be working properly to produce erroneous point correspondences.

Since both the calibration and the synchronization are the most crucial requirements for 3D video production, it is highly recommended to check the calibration errors before optimizing the parameters.

2.3.1.2.3 Bundle Adjustment

By the previous two steps, all calibration parameters have been estimated. One standard metric to evaluate the accuracy of the estimated parameters is the *reprojection error*. That is, for each corresponding point pair p_i^k and p_j^k of camera C_i and C_j , compute the 3D point P^k from them by triangulation, and reproject P^k onto the image planes again. Let \check{p}_i^k and \check{p}_j^k be the reprojection of P^k on the image planes of cameras C_i and C_j , respectively. Then the reprojection error is defined by

$$E(C_i, C_j) = \sum_k \{|p_i^k - \check{p}_i^k|^2 + |p_j^k - \check{p}_j^k|^2\}. \quad (2.11)$$

The goal of the non-linear optimization is to minimize this error for all cameras. That is, it optimizes a set of parameters which minimizes

$$E = \sum_{C_i \neq C_j \in C} E(C_i, C_j), \quad (2.12)$$

where C is the set of cameras. This optimization is called the *bundle adjustment*; it optimizes the calibration parameters by adjusting the bundle of light rays from each camera center to its image feature points so that corresponding rays from multiple cameras intersect each other in the 3D space.

In practice this non-linear optimization is done by Levenberg–Marquardt algorithm. Furthermore the sparse implementation of Levenberg–Marquardt algorithm can perform better since the Jacobian of Eq. (2.12) is significantly sparse. In addition, as pointed out by Hernandez et al. [11], modifying camera position T has very similar computational effects to shifting image center (u_0, v_0) in particular for circular camera arrangements, and hence fixing (u_0, v_0) through the optimization can perform better.

One important point in implementing the extrinsic parameter estimation is the estimation method of P^k from p_i^k and p_j^k . As discussed in [10], it is not a good idea to estimate P^k by the midpoint of the common perpendicular to the two rays through p_i^k and p_j^k , since it is not projective-invariant. Instead, [10] suggested to use linear triangulation methods or to solve a 6-degree polynomial.

2.3.1.2.4 Global Scale and Position Adjustment

The last step of the geometric calibration is to transform the world coordinate system used for the extrinsic parameter estimation into a physical one: determine the scale parameter of the common coordinate system to which all camera coordinate systems were transformed in the extrinsic parameter estimation. One simple practical method for it is to measure three points p_o , p_x , and p_y on the studio floor. p_o defines the origin of the physical coordinate system, the directions from p_o to p_x and p_y defines X - and Y -axes, respectively. The Z -direction is given by the cross

Fig. 2.9 Global scale and position adjustment using a chessboard on the floor



product of the X - and Y -directions. For example, place a chessboard designed with physical measures on the floor (Fig. 2.9). Let $\{R_i, T_i\}$ ($i = 1, \dots, N$) (N : number of cameras) denote the optimal extrinsic camera parameters obtained by the bundle adjustment. Then, select two cameras i' and i'' ($i' \neq i''$) which can best observe the chessboard and apply Zhang's method [33] to estimate the rotation and translation w.r.t. the floor as \hat{R}_j and \hat{T}_j ($j = i', i''$). The global scale parameter is given by comparing the distance between camera i' and i'' in the two different coordinate systems. That is,

$$\lambda = \frac{|\hat{T}_{i'} - \hat{T}_{i''}|}{|T_{i'} - T_{i''}|} \quad (2.13)$$

is the global scale parameter to be applied for the result of the bundle adjustment. Finally, in order to describe the camera positions and postures w.r.t. the floor, $\{R_i, T_i\}$ ($i = 1, \dots, N$) should be transformed to

$$\begin{aligned} R'_i &= \hat{R}_{i'} R_i^\top R_i \\ T'_i &= \lambda (\hat{R}_{i'} R_i^\top T_i - \hat{R}_{i'} R_{i'}^\top T_{i'}) + \hat{T}_{i'}, \end{aligned} \quad (2.14)$$

which represent the positions and postures of cameras in the physical coordinate system. With this representation, we can easily design object actions in the 3D video studio.

Note that the calibration accuracy of the above process does not affect the reconstruction accuracy of 3D object because it uniformly transforms all camera coordinate systems by a rotation and a translation. The accuracy of camera calibration in each 3D video studio we developed will be shown later.

2.3.1.3 Active Camera Calibration

While all geometric parameters of static cameras are fixed, those of active cameras can be dynamically changed during video capturing. Typical controllable parameters of active cameras include *pan*, *tilt*, *dolly*, and *zoom*. While pan, tilt, and dolly controls modify only the position of the projection center geometrically, zooming changes all camera parameters including the focal length, the projection center, the lens distortion, and the image resolution, since the zoom control modifies the entire optical system configuration of a camera.

Thus from a viewpoint of camera calibration, active cameras without zooming are a reasonable class of practically usable active cameras; the camera calibration process is required to estimate the position of the projection center dynamically while the other parameters are kept fixed.

In [30], we developed the *fixed-viewpoint pan-tilt camera*, where (1) the pan and tilt axes intersect with each other and (2) the projection center is aligned at the intersecting point. With this camera, the projection center is fixed during any pan-tilt controls, and hence it can be calibrated just as a static camera, which greatly facilitates the development of active object tracking systems to monitor 3D motion trajectories of objects [17] as well as high-resolution panoramic image capture systems.

One important technical problem when employing active cameras is the synchronization between the camera control and the image capture. That is, since these two processes usually run asynchronously, some synchronization mechanisms should be introduced to associate the viewing direction of a camera with a captured image. In [16], we proposed the *dynamic memory architecture* to virtually synchronize asynchronous processes. With this mechanism, each captured video frame can be annotated by synchronized pan and tilt parameter values. Note that pan and tilt values obtained from the camera controller are not accurate enough to be used as calibration parameters and hence the ordinary camera calibration should be done using them as initial estimates.

The calibration of active cameras, except for the fixed-viewpoint pan-tilt camera, involves many difficult technical problems including the camera model itself and hence its accuracy is limited. We will discuss them in Chap. 3 in detail.

2.3.2 Photometric Calibration

A camera records light flux converging to its projection center as a 2D array of pixel intensity values. While the geometric calibration models geometric aspects of this imaging process, the light flux has photometric characteristics such as colors (i.e. wave length of light) and powers (i.e. irradiance), which are also transformed through the imaging process. The goal of the photometric calibration is to rectify the photometric transformations by a camera.

Here, we consider the following two practical characteristics for the photometric calibration.

Gain: The *gain* defines the transformation from incident light intensities to image pixel values. First of all, to use cameras as physical sensors, the *γ correction* should be done to make this transformation linear; most cameras transform incident light intensities nonlinearly to image pixel values to make captured images look natural on displays or printed papers.

Since ordinary color cameras employ the RGB decomposition of incident light to record RGB image intensity values for each pixel, the gain is defined for each color channel. Then, the adjustment of RGB gains, which is called *color balance*

or *white balance*, should be done to capture high fidelity color images. Moreover, image sensor sensitivity and electronic circuit characteristics vary from camera to camera even if they are of the same type, making color calibration of multi-camera systems much harder.

Vignetting: Ordinary lens systems introduce *vignetting*: central areas of an image become brighter than peripheral areas. That is, the latter can receive less light rays compared to the former due to (1) multiple optical elements in a lens system (optical vignetting) and (2) the angle of incoming light (natural vignetting by the cosine fourth law). Compared to color calibration, vignetting elimination is rather easy if lens parameters are not dynamically changed.

In multi-camera systems, each camera observes a different part of the scene from a different viewpoint. This means that lighting environments vary from camera to camera. To calibrate lighting environments in a 3D video studio, 3D distributions of light sources and inter-reflections in the studio have to be modeled. These will be discussed in Chap. 6.

In this section, we assume we can prepare uniform lighting environments for the photometric calibration and present two practical photometric calibration methods for multi-camera systems: relative and absolute methods. The former normalizes photometric characteristics to be shared by all cameras, while the latter establishes their transformations to standard ones defined by reference data.

2.3.2.1 Relative Multi-camera Photometric Calibration

A standard idea of gain and vignetting correction is to measure a specified point in the scene by different pixels of an image sensor by moving a camera. That is, align the point at central and then peripheral image pixels one by one, and estimate parameters of a vignetting model. Kim and Pollefeys [14] proposed a method which estimates vignetting parameters from overlapped image areas in a patch-worked panoramic image. This method suits well for mobile camera and can calibrate spatial gain bias and vignetting of single-camera systems.

For multi-camera systems, we proposed an idea of object-oriented color calibration in [21]. The idea is to optimize vignetting and gain parameters of cameras to minimize observed color differences of a specified 3D object surface. The following process is applied to each color channel independently.

Let p denote an identifiable point on the 3D object surface and p_{C_i} the pixel representing the projection of p on the camera C_i image plane. Then, the ideal intensity value l at p_{C_i} is transformed first by a simplified Kang-and-Weiss model [34] representing the lens vignetting:

$$l' = \frac{1 - ar}{(1 + (r/f)^2)^2} l, \quad (2.15)$$

where r denotes the distance from the image center (u_0, v_0) to p . f and a denote the vignetting parameters. Then the intensity is transformed by the gain adjustment

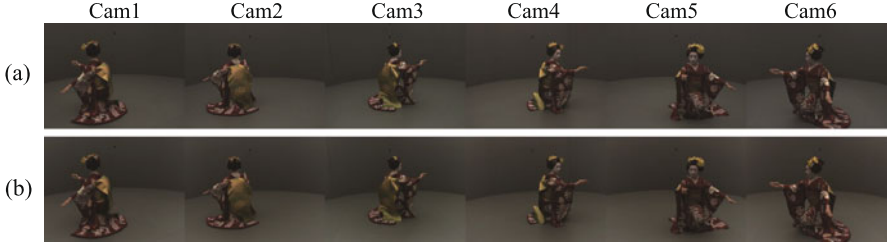


Fig. 2.10 (a) Originally captured multi-view images, (b) photometrically calibrated multi-view images. ©2009 IPSJ [22]

process as follows, assuming the γ correction has been done already:

$$l'' = \alpha l' + \beta, \quad (2.16)$$

where α and β denote the scale and bias factors. Reversing these transformations, the ideal intensity can be estimated from the observed intensity:

$$l = F(l'') = \frac{(l'' - \beta)(1 + (r/f)^2)^2}{\alpha(1 - ar)}. \quad (2.17)$$

Then, the goodness of the gain and vignetting parameters for p can be evaluated by

$$E(p) = \text{VAR}\{F_{C_i}(I_{C_i}(p_{C_i}))\}, \quad (2.18)$$

where C_i denotes a camera which can observe p without occlusion, $I_{C_i}(p_{C_i})$ the observed intensity of p_{C_i} , F_{C_i} the function defined in Eq. (2.17) for C_i , and $\text{VAR}\{\cdot\}$ the function to compute the variance. Note that p should be on a Lambertian surface because its radiance should be independent of viewing angles of C_i s.

Let P denote a set of Lambertian surface points. Then, apply Levenberg–Marquardt method to estimate the optimal gain and vignetting parameters which minimize the following objective function.

$$E = \sum_{p \in P} E(p). \quad (2.19)$$

Figure 2.10 shows the result of the photometric calibration of multi-view images. Figure 2.11 demonstrates that photometric characteristic variations of uncalibrated cameras can introduce visible artifacts in images rendered from 3D video. Here the simplest view-independent texture generation method in Sect. 5.3 is used to demonstrate the color differences across original images.

Notice that the relative photometric calibration normalizes photometric characteristics of multi-view cameras so that multi-view observations of a 3D surface point give the same pixel intensity value. Hence it does not guarantee that the calibrated color is the “true” color of the object.

Fig. 2.11 Textures generated from Fig. 2.10(a) and (b), respectively. The red arrows indicate texture boundaries introduced by photometric characteristics variations among the cameras. ©2009 IPSJ [22].

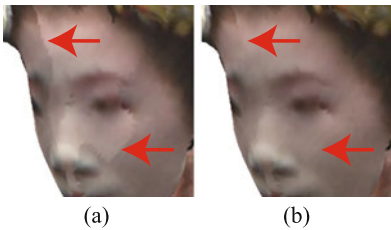


Fig. 2.12 Macbeth color checker. The triplet of hexadecimal values attached to each color patch denotes approximated 8-bit RGB values [24]

#6f4f38	#ceaa99	#5e8fb8	#607c43	#9b97d2	#8ddad6
#d98404	#3764b9	#d55f72	#6d2181	#bcd448	#f0bb30
#004291	#45a53f	#d11b3f	#ecd621	#dc41b7	#00acc1
#e8e8e8	#d0ced9	#abb0b4	#7e8a8a	#4e5c5c	#273333

2.3.2.2 Absolute Multi-camera Photometric Calibration

Assuming that the vignetting calibration is done, the absolute color calibration adjusts RGB color channel gains of a camera so that RGB values for reference color patterns coincide with predefined standard responses. Figure 2.12 shows a well-known color pattern called *Macbeth color checker*, where each color patch is associated with predefined standard RGB values [24]. The color calibration with a standard color pattern also requires standard lighting environments: the pattern should be uniformly lit by a standard light source such as defined by ISO/IEC standards.

As is well known, since RGB values denote spectral integrals, the accuracy of the above-mentioned RGB-based color calibration is limited. Thus, physics-based color calibration should be employed to attain the truly absolute color calibration: estimate spectral filtering characteristics of RGB channels from a reference pattern and a light source whose spectral radiance and radiant characteristics are known, respectively.

To evaluate the usability of standard color samples, such as Munsell standard colors, in the physics-based color calibration, we measured spectral characteristics of radiance intensities of 1,016 color samples lit by a standard light, where spectral characteristics of each color sample is represented by 176 radiance intensity values from 380 nm to 730 nm with 2 nm sampling pitch. Then, we computed the major principal components. Table 2.4 shows eigen values and residual errors for 16 major principal components. From these results, we can observe that spectral characteristics of Munsell color samples can only be represented by several major spectral bases. This implies that detailed spectral characteristics of cameras and lighting en-

Table 2.4 Dimensionality reduction of Macbeth colors by PCA

# of principal component	Eigenvalue	Approx. error
0		100.000
1	2.1544e−02	18.011
2	5.1743e−04	9.592
3	1.3787e−04	5.486
4	4.3802e−05	3.228
5	1.1529e−05	2.290
6	4.7269e−06	1.766
7	3.1202e−06	1.311
8	1.7696e−06	0.961
9	7.4854e−07	0.766
10	4.4186e−07	0.624
11	2.6615e−07	0.519
12	1.9256e−07	0.428
13	1.6722e−07	0.329
14	8.3086e−08	0.269
15	5.5517e−08	0.218
16	3.2762e−08	0.182



vironments cannot be estimated with such color samples; the dimension of spectral characteristic space is degenerated.

To estimate spectral characteristics of cameras, we need to utilize additional reference measurements given by special optical systems such as spectrometer [19], multi-spectral camera [18], and hyper-spectral sensor [7]. These techniques play an important role on digital archiving of cultural assets such as ancient tapestries, statues, etc. In addition, knowledge about spectral characteristics of reference objects can help to calibrate such sensors. ISO/TR 16066:2003 [12] provides spectral color data of more than 50 thousand common objects as well as their reflectance and transmittance characteristics in order to calibrate spectral response of image sensors.

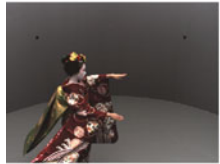







While the absolute photometric calibration can be conducted for each camera independently before installation in a 3D video studio, lighting environments of the studio should be estimated to obtain calibrated RGB values. As will be discussed in Chap. 6, the lighting environment estimation itself involves difficult problems. Especially, it would be almost impossible to estimate the 3D spatial distribution of detailed spectral characteristics of lighting environments, because an object in motion disturbs lighting environments by its shadows as well as inter-reflections with the background scene.

In summary, it would be a practical method for multi-camera photometric calibration to employ the relative multi-camera photometric calibration and then normalize RGB values based on the RGB responses of an absolutely calibrated camera.

Studio A

	
<i>Walking</i> 150frames@12.5fps Active tracking	<i>Radio-controlled animal toy</i> 820frames@12.5fps Active tracking

Studio B

			
<i>Red MAIKO</i> 9000frames@25fps Complex and non-rigid shape	<i>Red MAIKO pair</i> 9000frames@25fps Complex and non-rigid shape	<i>SAMURAI-1</i> 500frames@25fps Complex and non-rigid shape	<i>SAMURAI-2</i> 500frames@25fps Complex and non-rigid shape
			
<i>Green MAIKO</i> 1frame Complex shape	<i>Capoeira</i> 2000frames@25fps Complex motion	<i>Yoga</i> 7500frames@25fps Complex motion	<i>Juggling</i> 1250frames@25fps Multiple objects

Studio C



	
<i>Gazing</i> 125frames@25fps Gazing at designated points	<i>Tony</i> 250frames@25fps Complex motion

Fig. 2.13 Multi-view videos captured by the three studios. *Each line* of the subtitles shows the title, captured length, and feature, respectively

2.4 Performance Evaluation of 3D Video Studios

Figure 2.13 shows multi-view videos captured by the three studios in Kyoto University described in Sect. 2.2.7. Each of them has different features such as active tracking, complex and non-rigid object shape, complex motion, etc. They will be used as input for our 3D video production algorithm described in the following chapters.

Table 2.5 Performances of the three studios described in Sect. 2.2.7

	Studio A	Studio B	Studio C
Capture space (m)			
Cylinder diameter	3.0	3.0	4.0
Cylinder height	2.0	2.2	2.2
Effective resolution (mm/pix)	3.9	2.0	2.0
Calibration accuracy (mm)	4.3	2.4	3.4

Table 2.5 reports the performance measures achieved in the three studios. The capture spaces are approximated by cylinders where the requirements for 3D video production are satisfied. In the table, the diameters and heights of the cylinders are described. As in most of 3D video studios, the object movable space is very limited to guarantee the high object surface observability.

The effective resolution implies the average physical distance between two neighboring pixels at the center of the capture space. The calibration accuracy is computed as the average 3D distance between a pair of rays from a pair of corresponding points in different views. The accuracy in 2D, that is, the reprojection errors of corresponding points are all in sub-pixel level.

The lower resolution and accuracy of Studio A can be ascribed to its lower camera resolution (VGA); Studio A was developed for tracking and multi-view object observation with pan/tilt/zoom active cameras. Studio C, on the other hand, was designed to realize a wider object movable space with the almost same number of cameras as Studio B. To this end, the field of view was increased by employing a larger imager (1/1.8 inch) as well as improving the camera resolution (UXGA). With these designs, the effective resolution of Studio C attained the same level as that of Studio B, while the calibration accuracy was degraded due to its enlarged capture area.

In summary, to enlarge the capture space as well as improve the effective resolution and calibration accuracy, we need to increase the number of cameras or employ active pan/tilt/zoom cameras. This problem is discussed in the next chapter.

2.5 Conclusion

This chapter discussed the design factors of multi-camera systems for 3D video studios and introduced our three implementations. While advanced imaging device, computer, and computer vision technologies make it rather easy to implement 3D video studios, many problems are left in (1) camera selection and arrangements to guarantee multi-view observability of an object in motion, (2) geometric and photometric camera calibrations to realize the “seamless” integration of multi-view video data, and (3) design and calibration of lighting environments. These are crucial requirements for successful 3D video production for Part II.

As noted at the beginning of this chapter, we designed and implemented 3D video studios with off-the-shelf cameras and lenses. Specially developed cameras

such as 4K and 8K cameras with professional lenses will improve the performance measures of 3D video studios shown in Table 2.5, while algorithms and technologies to solve the problems (1) and (2) above are left for future studies.

The second generation of 3D video studios will employ new imaging sensors such as time-of-flight cameras or active-stereo systems to directly obtain 2.5D video data. Their calibration, synchronization, and data integration with ordinary video cameras will require the development of new technologies. Similarly, it would be another interesting augmentation of 3D video studios to introduce audio capturing devices such as microphone arrays for recording 3D acoustic environments. To integrate 3D visual and acoustic scenes, cross-media synchronization and calibration methods should be developed.

References

1. Batlle, J., Mouaddib, E., Salvi, J.: Recent progress in coded structured light as a technique to solve the correspondence problem: a survey. *Pattern Recognit.* **31**(7), 963–982 (1998)
2. Bayer, B.E.: US Patent 3971065: Color imaging array (1976)
3. Bouguet, J.-Y.: Camera Calibration Toolbox for Matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/
4. Bradski, G.: The OpenCV Library (2000). <http://opencv.willowgarage.com>
5. Virtualizing Engine. Private communication with Profs. Takeo Kanade and Yaser Sheikh, Robotics Institute, Carnegie Mellon University, PA (2011)
6. Fitzgibbon, A.W., Zisserman, A.: Automatic camera recovery for closed or open image sequences. In: *Proc. of European Conference on Computer Vision*, pp. 311–326 (1998)
7. Gevers, T., Stokman, H.M.G., van de Weijer, J.: Colour constancy from hyper-spectral data. In: *Proc. of British Machine Vision Conference* (2000)
8. Goldlücke, B., Magnor, M.: Joint 3D-reconstruction and background separation in multiple views using graph cuts. In: *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 683–688 (2003)
9. Guillemaut, J.Y., Hilton, A., Starck, J., Kilner, J., Grau, O.: A Bayesian framework for simultaneous matting and 3D reconstruction. In: *Proc. of International Conference on 3-D Digital Imaging and Modeling*, pp. 167–176 (2007)
10. Hartley, R.I., Zisserman, A.: *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge (2000)
11. Hernandez, C., Schmitt, F., Cipolla, R.: Silhouette coherence for camera calibration under circular motion. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(2), 343–349 (2007)
12. ISO/TR 16066: Standard Object Colour Spectra Database for Colour Reproduction Evaluation (SOCS) (2003)
13. Ivanov, Y., Bobick, A., Liu, J.: Fast lighting independent background subtraction. *Int. J. Comput. Vis.* **37**(2), 199–207 (2000)
14. Kim, S.J., Pollefeys, M.: Robust radiometric calibration and vignetting correction. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(4), 562–576 (2008)
15. Levoy, M., Hanrahan, P.: Light field rendering. In: *Proc. of ACM SIGGRAPH*, pp. 31–42 (1996)
16. Matsuyama, T., Hiura, S., Wada, T., Murase, K., Toshioka, A.: Dynamic memory: architecture for real time integration of visual perception, camera action, and network communication. In: *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 728–735 (2000)
17. Matsuyama, T., Ukita, N.: Real-time multitarget tracking by a cooperative distributed vision system. *Proc. IEEE* **90**(7), 1136–1150 (2002)

18. Miyake, Y., Yokoyama, Y., Tsumura, N., Haneishi, H., Miyata, K., Hayashi, J.: Development of multiband color imaging systems for recordings of art paintings. In: Proc. of SPIE, pp. 218–225 (1998)
19. Morimoto, T., Mihashi, T., Ikeuchi, K.: Color restoration method based on spectral information using normalized cut. *Int. J. Autom. Comput.* **5**, 226–233 (2008)
20. Nister, D.: An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(6), 756–770 (2004)
21. Nobuhara, S., Kimura, Y., Matsuyama, T.: Object-oriented color calibration of multi-viewpoint cameras in sparse and convergent arrangement. *IPSP Trans. Comput. Vis. Appl.* **2**, 132–144 (2010)
22. Nobuhara, S., Tsuda, Y., Ohama, I., Matsuyama, T.: Multi-viewpoint silhouette extraction with 3D context-aware error detection, correction, and shadow suppression. *IPSP Trans. Comput. Vis. Appl.* **1**, 242–259 (2009)
23. Okutomi, M., Kanade, T.: A multiple-baseline stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* **15**(1), 353–363 (1993)
24. Pascale, D.: RGB coordinates of the ColorChecker (2006). http://www.babelcolor.com/main_level/ColorChecker.htm
25. PMDTechnologies GmbH: CamCube3.0 (2010)
26. Salvi, J., Pagès, J., Batlle, J.: Pattern codification strategies in structured light systems. *Pattern Recognit.* **37**(4), 827–849 (2004)
27. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **47**, 7–42 (2002)
28. Starck, J., Maki, A., Nobuhara, S., Hilton, A., Matsuyama, T.: The multiple-camera 3-d production studio. *IEEE Trans. Circuits Syst. Video Technol.* **19**(6), 856–869 (2009)
29. Toyoura, M., Iiyama, M., Kakusho, K., Minoh, M.: Silhouette extraction with random pattern backgrounds for the volume intersection method. In: Proc. of International Conference on 3-D Digital Imaging and Modeling, pp. 225–232 (2007)
30. Wada, T., Matsuyama, T.: Appearance sphere: Background model for pan-tilt-zoom camera. In: Proc. of International Conference on Pattern Recognition, pp. A-718–A-722 (1996)
31. Yamaguchi, T., Wilburn, B., Ofek, E.: Video-based modeling of dynamic hair. In: Proc. of PSIVT, pp. 585–596 (2009)
32. Zeng, G., Quan, L.: Silhouette extraction from multiple images of an unknown background. In: Proc. of Asian Conference on Computer Vision, pp. 628–633 (2004)
33. Zhang, Z.: A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(11), 1330–1334 (2000)
34. Zheng, Y., Yu, J., Kang, S., Lin, S., Kambhamettu, C.: Single-image vignetting correction using radial gradient symmetry. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)

3D Video and Its Applications

Matsuyama, T.; Nobuhara, S.; Takai, T.; Tung, T.

2012, XV, 346 p., Hardcover

ISBN: 978-1-4471-4119-8