Multiview acquisition systems

Frédéric Devernay, Yves Pupulin, Yannick Rémion

To cite this version:

Frédéric Devernay, Yves Pupulin, Yannick Rémion. Multiview acquisition systems. Laurent Lucas and Céline Loscos and Yannick Remion. 3D Video: From Capture to Transmission, Wiley-ISTE, 2013, Computer Engineering Series, 978-1-84821-507-8. <hal-00856832>

HAL Id: hal-00856832

https://hal.inria.fr/hal-00856832

Submitted on 2 Sep 2013

Chapter 3

Multiview acquisition systems

Chapter by Frédéric Devernay, Yves Pupulin and Yannick Remion, from the book “3D Video: From Capture to Transmission”, Wiley-ISTE, 2013, Laurent Lucas, Yannick Rémion, Céline Loscos (eds), ISBN: 978-1-84821-507-8.

3.1 Introduction: what is a multiview acquisition system?

Multiview acquisition, the focus of this chapter, relates to the capture of synchronized video data representing different viewpoints of a single scene. In contrast to video surveillance systems, which deploy multiple cameras to visually cover a large scale environment to be monitored with little redundancy, the materials, devices or systems used in multiview acquisition are designed to cover several perspectives of a single, often fairly restricted, physical space and use redundancy in images for specific aims:

- for 3D stereoscopic or multiscopic visualization of captured videos;
- for real scene reconstruction/ virtualization :
  - 2.5D reconstruction of a depth map from a given viewpoint;
  - textured 3D reconstruction of digital models, avatars of real objects,
  - motion capture for realistic animation of virtual actors;
- for various and complementary adjustments in control room or during post production:
  - “mosaicking” views providing a panoramic view or a high resolution image,
  - a virtual camera moving at frozen time or very slowly (bullet time),
  - mixing the real/ virtual (augmented reality - AR)
  - view interpolation (free viewpoint TV - FTV),
  - focus modification after shooting (refocus),
  - increasing video dynamics (high dynamic range - HDR),
  - etc.

Depending on the final application, the number, layout and settings of cameras can fluctuate greatly. The most common configurations available today include:

- “Binocular systems” yielding two views from close-together viewpoints; these systems are compatible with 3D stereoscopic visualization (generally requiring glasses) and depth reconstruction with associated post production methods (AR, FTV);

- Lateral or directional multiview systems [1] . provide multiple views from close-together viewpoints (generally regularly spaced), each placed on the same side of a scene. These systems produce media adapted to autostereoscopic 3D visualization, “frozen time” effects within a limited range, and a depth reconstruction or more robust “directional” 3D reconstruction than is the case of binocular reconstruction with the same post production techniques (AR, FTV). The multiplication of different perspectives also allows using different settings for each camera which, with the strong redundancy in capture, renders other post production methods possible (refocus or HDR, for example);

- Global or omnidirectional multiview systems [1] deploy their multiple viewpoints around of the target space. These systems are principally designed for bullet time in a wide angular motion, 3D reconstruction and motion capture (MoCap).

Alongside these purely video-based solutions, hybrid systems adding depth sensors (“Z-cams”) to video sensors are also interesting. The captured depth can theoretically provide direct access to the majority of desired post-productions. The number of video sensors as well as depth sensor’ resolution and spatial limitations can, however, restrict some of these post production processes. These hybrid systems, however, will not be examined within this book.

All these materials share the need to synchronize and calibrate (often even with geometric and/ or colorimetric corrections) information captured by different cameras or Z-cam and often have different accompanying capabilities regarding:

- recording signals from all sensors without loss of data;

- processing all data in real-time, which demands a significant computation infrastructure (often using distributed computing).

This chapter will introduce the main configurations cited above in a purely video multiview capture context, using notable practical examples and their use. Each time, we will also propose links to databases providing access to media produced by devices within each category.

3.2 Binocular systems

3.2.1 Technical description

Capturing binocular video, also known as stereoscopy or, more recently "‘3D stereoscopy"’ (3DS), requires the use of two cameras [2] connected by a rigid or articulated mechanical device known as a "‘stereoscopic rig"’. The images taken can be projected either on a stereoscopic display device (such as a cinema screen or a 3D television, most commonly) [DB10], or used to extract the scene’s 3D geometry, in the form of a depth map, using stereo correspondence algorithms.

[1] Term used within this book.

[2] In photography, where the scene is fixed, we only need a single device that is moved along a slider between the left and right views.

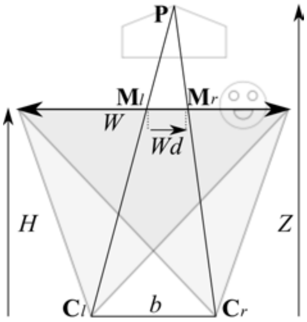
The shooting geometry

Filming is carried out using two cameras with the same optical parameters (focal length, focus distance, exposure time, etc.), pointing roughly in the same direction, orthogonal to the line connecting their optical centers (which is known as the *baseline*). The optical axes can be parallel or convergent.

Ideally, to simplify stereoscopic correspondence, the two optical axes must be strictly parallel, orthogonal to the baseline, and the two image planes must be identical. In this situation, the corresponding points have the same y-coordinate in both images. However, if the cameras are convergent (i.e. the optical axes converge at a finite distance) or if the alignment is approximate, the images taken by the camera can be rectified (see section 5.4) to get back to the ideal situation. Rectification is therefore an important post-production phase for stereoscopic films (see section 3.2.2).

The main geometric parameters for stereoscopic recording and stereoscopic visualization are described in figure 3.1. b, W and H are the parameters of the stereoscopic camera and Z is the distance from a 3D point to the plane passing through the stereoscopic baseline and parallel to the image planes. The triangles $\mathbf{M_lPM_r}$ and $\mathbf{C_lPC_r}$ are homothetic. As a result : $(Z - H)/Z = dW/b$. This allows us to simply express the relations between the stereoscopic disparity d, expressed as a fraction of the image's width W and the distance Z, similar to that shown in chapter 7 :

$$d = \frac{b}{W}\frac{Z-H}{Z}, \quad \text{or} \quad Z = \frac{H}{1-dW/b} \tag{3.1}$$



Symbol	Camera	Display
$\mathbf{C}_l$, $\mathbf{C}_r$	optical center	optical center (eyes)
$\mathbf{P}$	point in the scene	perceived 3D point
$\mathbf{M}_l$, $\mathbf{M}_r$	points in the image in $\mathbf{P}$	points on the screen
b	interocular	interocular (eyes)
H	convergence distance	distance from the screen
W	width of the convergence plane	size of the screen
Z	real depth	perceived depth
d	right-left disparity (as a fraction of W)	

Figure 3.1: Geometry of the stereoscopic shooting device and that of the stereoscopic display device can be described by the same low number of parameters

Perceived geometric distortions

If stereoscopic video is designed to be projected on a stereoscopic display device whose parameters are b', W' and H', the depth Z' perceived by stereoscopy[3] can be calculated according to the disparity d (equation (3.2)). By eliminating the disparity d from (3.1) and (3.2), in (3.3) we obtain the relation between the real depth Z and the perceived depth Z' which will be applied to the multiscopic example in chapter 4:

[3] Stereoscopy is combined with a number of other monocular indices to create the 3D perception of the scene[Lip82]: light and shade, relative size, interposition, texture gradient, aerial perspective, perspective, flow, etc.

$$Z' = \frac{H'}{1 - dW'/b'} \tag{3.2}$$

$$Z' = \frac{H'}{1 - \frac{W'}{b'}(\frac{b}{W}\frac{Z-H}{Z})} \quad \text{or} \quad Z = \frac{H}{1 - \frac{W}{b}(\frac{b'}{W'}\frac{Z'-H'}{Z'})} \tag{3.3}$$

There is ocular divergence when $Z' < 0$ ($d' > \frac{b'}{W'}$), i.e. when the on-screen binocular disparity is larger than the viewer's interocular. In general, real objects that are very far away ($Z \to +\infty$) are perceived at a finite distance or create divergence, depending on whether $\frac{W'}{b'}\frac{b}{W}$ is smaller or greater than 1. We consider than an ocular divergence in the order of 0.5° is acceptable for short durations, and this trick is used by stereographers to artificially augment the depth available behind the movie screen.

In the case of 3D television, the disparity limits due to the conflict between convergence and accommodation [ENO05, UH07, YEM04] render large (either positive or negative) disparities uncomfortable. The depth of focus of the human eye is of the order of around 0.3 δ (diopters) in normal situations[4], which, on a screen placed 3 meters away, gives a depth of focus ranging from $1/(\frac{1}{3} + 0.3) \approx 1.6$ m to $1/(\frac{1}{3} - 0.3) = 30$ m. In practice, TV production rules are much stricter. 3DTV programs are produced with disparities ranging from -1 % to $+2$ % of the screen width [5] to remain in this comfort zone[6], with disparities temporarily ranging from -2.5 % to $+4$ %, which completely prevents reaching the divergence limit on private projection devices.

We can see also that the situation where the perceived depth is strictly identical to the real depth ($Z' = Z$) can only be obtained if all parameters are equal, which is known as the "‘orthostereoscopic"’ configuration (this configuration is often used for IMAX 3D films since the geometry of the projection device is known beforehand).

For a 3D fronto-parallel plane placed at a distance Z, we can calculate the scale factor s between the distances measured within this frame and the distances in the convergence plane: $s = H/Z$. We can also calculate the image scale factor σ' which explains the extent to which an object placed at a depth of Z or the disparity d is perceived as being enlarged ($\sigma' > 1$) or reduced ($\sigma' < 1$) in the directions X and Y with respect to objects in the convergence plane ($Z = H$):

$$\sigma' = \frac{s'}{s} = \frac{H'}{Z'}\frac{Z}{H} = \frac{1 - dW'/b'}{1 - dW/b} \tag{3.4}$$

Of course, for objects in the screen plane ($d = 0$), we have $\sigma' = 1$. The relation between Z and Z' is linear if and only if $W/b = W'/b'$, in which case $\sigma' = 1$ and $Z' = ZH'/H$. We refer to this configuration as being "‘orthoplastic"’ configuration (an orthostereoscopic configuration is, above all, orthoplastic).

A small object with a width of ∂X and a depth of ∂Z, placed at Z is perceived as an object with the dimensions $\partial X' \times \partial Z'$ at a depth of Z', and the *roundness factor* ρ measures how much the object's proportions are modified:

$$\rho = \frac{\partial Z'}{\partial Z} / \frac{\partial X'}{\partial X} = \frac{\partial Z'}{\partial Z} / \frac{W'/s'}{W/s} = \sigma' \frac{W}{W'}\frac{\partial Z'}{\partial Z} \tag{3.5}$$

[4] More precise studies [MMN99] have shown that this also depends on parameters such a pupil diameter, wavelength and spectral composition.

[5] Negative disparities correspond to points closer to the screen and positive disparities to disparities further away.

[6] See, for example, the production guidelines of Sky 3D in the UK: www.sky.com/shop/tv/3d/producing3d.

In the screen's frame($Z = H$ and $Z' = H'$), the roundness factor can be simplified as:

$$\rho_{\text{écran}} = \frac{W}{W'} \frac{\partial Z'}{\partial Z}_{(Z=H)} = \frac{b}{H} \frac{H'}{b'} \tag{3.6}$$

A roundness factor equal to 1 indicates that a sphere is perceived exactly as a sphere, a smaller roundness factor indicates that it is perceived as a sphere flattened in the depth direction and a larger roundness factor indicates that it is perceived as a ellipsoid stretched in the depth direction. The roundness of an object in the screen plane is equal to 1 if, and only if, $b'/b = H'/H$. In order for this to be the case in the whole space, it is necessary that $b'/b = W'/W = H'/H$. As a result, the only geometric configurations which preserve roundness everywhere are identical to the display configuration up to a scale factor; these are "‘orthoplastic"’ configurations. Even if the geometry of the display device is known during filming, this imposes strict constraints on how the film is shot, which can be very difficult to follow in different situations (i.e. when filming sports events or wildlife documentaries). On the other hand, since the viewer's interocular b' is fixed, this indicates that a film can only be projected on a screen of a given size W' placed at a given distance H', which is in contradiction with the large variability of projection devices and movie theaters. We therefore refer to "‘hyperplastic"’ or "‘hypoplastic"’ configurations when the roundness is larger or smaller than 1 respectively. The roundness in the screen plane also increases when we move away from the screen and it is independent of screen size, which is counter-intuitive; the majority of viewers expect to perceive "‘more 3D"’ when approaching a large screen.

Another important point to make is that a film, shot to have a specific roundness for a cinema screen positioned on average 15m away, will see its roundness divided by five once projected on a 3D TV screen placed 3m away, which in part explains the current dissatisfaction of 3DTV viewers. This effect can be counter-balanced by specific post-production for media designed for private viewing (home cinema), e.g. for 3D Blu-ray, although there are few titles which benefit from this treatment. Of course, this reduction in roundness is, in part, compensated by monoscopic depth cues. Besides, the roundness used in 3D cinema films is, in reality, between 0.3 and 0.6, depending on the desired dramatic effect [Men09], in order to favor the viewer's visual comfort.

3.2.2 Principal uses

Cinema and 3D television

Cinema and television rigs are for the most part heavy systems which often use a semi-reflective mirror to obtain distances for the camera interocular shorter than the diameter of the lens [Men11] (see left in figure 3.2). A number of manufacturers today produce compact semi-professional integrated stereoscopic cameras but their field of use is reduced, notably due to the fact that the interocular of these cameras is generally fixed whilst stereoscopic filming requires an adequate tuning of all stereoscopic parameters; merely adding a second camera alongside the first is not enough for 3D-S filming.

Stereoscopy, a new and different art 2D cinema, in order to exist, has (i) had to study the function of the brain in order to trick it into believing that a series of fixed images are really showing movement, (ii) to survey, from experience gained from photography, techniques which enable this illusion and develop a complete cinematographic chain and (iii) to invent the parameters of a new art, which is the role of artists involved in the production of films, followed by engineers producing tools enabling these new artistic practices.

Stereoscopy is both a continuous evolution and a turning point in cinematography due to the fact that, as with photography, it must use current techniques and develop others. To do so, it is essential to:

- restudy the brain and the visual system and examine how to trick it, not only temporally but also spatially by recreating the illusion of a three dimensional space whilst, in reality, there are only two 2D images;
- improve recording and postproduction stereoscopy tools in the cinematographic chain and produce new ones based on cerebral observations in order to ensure that this new illusion is comfortable;
- enable the invention of a filming technique based on these different parameters which contribute to creating this illusion.

The cinematographic parameters on which traditional filming relies are well known. However, the rules that govern the stereoscopic parameters in order to create this new illusion have not yet been established. Based on the way the human visual system works, they should simulate (i) how convergence is, in general, coupled with accomodation, and (ii) 3D vision resulting from the distance between both eyes, a parameter which varies slightly throughout the lifespan of each individual and between individuals.

However, simply shooting with an interocular equal to the average interocular of a population sample cannot, contrary to some ophthalmological studies, be considered sufficient. Indeed, stereoscopy uses these two parameters (interocular and convergence) to create emotion and feeling, exactly as the lenses used on a camera do not try to reproduce human perspective vision but reform it depending on the medium used. If we push these variations in distance to the extreme, on the one hand we have the value 0, which corresponds to two identical 2D images and, on the other hand, interaxial distances without any relationship with the geometry of the human visual system. NASA, for example, has produced stereoscopic images of Earth with a distance of almost seventy meters between the two viewpoints.

To create a rig, the interocular distance must be able to varry from 0 to the greatest usable value for some kind of scene. In general, for a standard configuration for comedy, a variation from a few millimeters to several centimeters corresponds to 90% of needs for fiction based filming. As a result, rigs used for close ups have interocular ranges between 0 and 100mm. Lastly, for long distance shots of clouds, for example, the distance between the two cameras may even extend to several meters and the side-by-side rigs are often adapted to specific needs for a given shot.

Computer-assisted production Whilst the rules for re-creating a universe in 3D have been known since the 19th century, the possibility of stereoscopic filming using rigs is much more recent and involves the use of a computer to analyze video streams and correct any potential faults. Given the fact that no mechanical, optical or electronic device is perfect, it is imperative to correct the recorded images as precisely as possible with a 3D corrector, in real time for television and in post-production for cinema. This was enabled by the invention of digital images which can correct each pixel individually.

Robotized rigs A rig must use synchronized cameras and lenses with perfectly synchronized and calibrated zoom, point and diaphragm movements. The rig itself is robotized and contains motors which adjust distance and convergence in real-time, as well as yaw/pitch/roll adjusting plates used to converge the two optical axes (the optical axes must be concurrent). In some cases, rigs have been used with more than two cameras, as was the case for the French language film

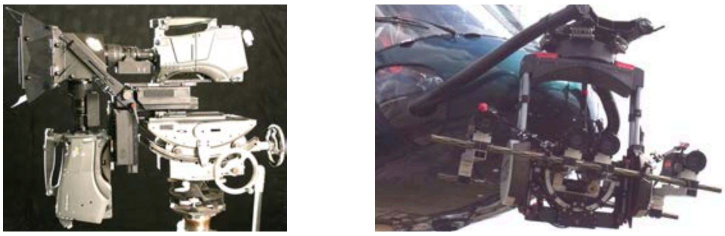


Figure 3.2: Examples of rigs: left, Binocle Brigger III in a studio configuration, a robotized rig for 3D TV, right, a heliborne rig with four cameras used by Binocle for the film La France entre ciel et mer

La France entre ciel et mer[France between sky and sea] which was filmed by Binocle with four cameras on a helicopter (see figure 3.2). In this case, the matching of four zooms and adjusting plates with four cameras demanded a huge degree of expertise since all optical centers had to be aligned as closely as possible. Examples of materials used to pilot the rig, and to directly control the geometric and photometric quality and faults include TaggerLive and TaggerMovie by Binocle[7], STAN – *Stereoscopic Analyzer* – by Fraunhofer HHI, SIP – *Stereoscopic Image Processor* – by 3ality Technica[8], the real time correction processor MPES-3D01 – often referred to as "‘3DBox"’ – by Sony, and Pure by Stereolabs[9].

Stereoscopic postproduction Postproduction tools have also been adapted to 3D cinema and algorithms specific to stereoscopy have been integrated into this software such as rectification, viewpoint interpolation and depth modifications, 2D to 3D conversion, color balancing of two streams, production of a depth map for 3D scene compositing, etc. These tools include the Ocula plugins suite for Nuke (The Foundry)[10], DisparityKiller (Binocle), and Mistika Post (SGO)[11].

Depth reconstruction

Binocular systems designed to produce a stereoscopic reconstruction of “partial” 3D data [12] are generally much simpler than those used for cinema or television. These are most often lightweight systems which are small, consume little energy and can be used by a vehicle or mobile robot, for example, and they almost always have a fixed interocular distance in order to simplify their calibration.

The majority of these systems use monochrome cameras, since with brightness alone is sufficient for stereoscopic correspondence, but color may bring additional functions such as the possibility of using color for segmentation tasks (such as skin color, for example) or object recognition. Cameras used in this kind of system generally use a single sensor, since the use of color (by the way of a Bayer matrix filter) results in a loss of spatial resolution in images and therefore affect the precision of reconstructed depth.

The choice of the optimal interocular distance value for reconstruction is a disputed subject but a simple rule of thumb can predict the final precision. The precision of the disparity d obtained by the stereoscopic correspondence algorithm can be presumed constant in the image

[7] www.binocle.com.

[8] www.3alitytechnica.com/3D-rigs/SIP.php.

[9] www.stereolabs.tv/products/pure/.

[10] www.thefoundry.co.uk/products/ocula/.

[11] www.sgo.es/mistika-post/.

[12] In the sense that they only contain the 3D information about the scene as seen from the stereo rig viewpoint.

(let us say 0.5 pixels). The error in the reconstructed depth Z is obtained by deriving equation (3.1): $\partial Z/\partial d = bHW/(b - dW)^2$, and $\partial Z/\partial d = Z^2W/(bH)$. The error increases with the square of the distance and theoretically decreases with the interocular distance b, so that theoretically the larger the interocular distance, the better the precision in depth reconstruction. However, when we increase the distance, stereoscopic matching between the images is more difficult and the precision of disparity d is strongly degraded when the b/H value increases. Experience shows that, as a rule of thumb, a b/H value between 0.1 and 0.3 represents a reasonable compromise between ease of stereoscopic correspondence and precision in depth reconstruction.

Any pair of rigidly linked and synchronized cameras can be used[13] to reconstruct depth using stereoscopic correspondence algorithms (the OpenCV software library provides calibration functions, stereoscopic correspondence and simple 3D reconstruction algorithms).

Commercial off-the-shelf systems are also available. They have the advantage of being solidly constructed, pre-calibrated or easy to calibrate, and sometimes propose optimized stereoscopic correspondence algorithms, using to the CPU or a dedicated FPGA (field-programmable gate array). Point Grey have developed the Bumblebee system[14] using two or three cameras with different sensors or focal length options and a SDK (software development kit) for computing depth maps on the CPU. The Tyzx DeepSea stereo vision system[15], proposed with several interocular distance options, uses a FPGA and a PowerPC CPU to compute disparity, and transmits the 3D data via ethernet.

Focus Robotics has developed nDepth[16], with a fixed interocular distance of 6cm, and a factory-calibrated monochrome sensor. Videre Design[17] has created stereo vision systems with fixed or variable interocular distances, with disparity computation carried out by the Small Vision System software (developed by SRI) or by a special chip (STOC- Stereo On Chip). Surveyor Corporation [18] sells the *Stereo Vision System* (SVS) which is a low cost solution for stereo with options such as embedded image capture, motorization and Wifi transmission, based on an open source firmware.

3.2.3 Related databases

The European QUALINET project[19] has collated and classified a number of multimedia databases with a specific section dedicated to 3D Visual Content Databases directing users towards databases of fixed images or multiview stereoscopic video. The MOBILE-3DTV project[20] also contains a number of reference stereoscopic sequences. Other high quality databases are also made available thanks to IEEE-3D *Quality Assesment Standard Group* [21] and the Sigmedia team at Trinity College Dublin[22].

[13] Synchronization is carried out either by a specific master-slave trigger connection between cameras or by the image transfer bus (for example, the majority of cameras manufactured by Point Grey are automatically synchronized when they are on the same "‘firewire"’ bus).

[14] www.ptgrey.com/products/stereo.asp.

[15] www.tyzx.com.

[16] www.focusrobotics.com/.

[17] http://users.rcn.com/mclaughl.dnai/.

[18] www.surveyor.com/.

[19] www.qualinet.eu, dbq.multimediatech.cz.

[20] www.focusrobotics.com/.

[21] http://grouper.ieee.org/groups/3dhf, ftp://165.132.126.47.

[22] www.tchpc.tcd.ie/stereo_database/.

3.3 Lateral or directional multiview systems

3.3.1 Technical description

This section examines systems and devices with close-together multiview (relative to the scene being filmed) sensors, often distributed evenly along a curve (whether rectilinear or not) or on a grid (flat or not). There are thus systems designed by mechanical assembly (linear or matricial) and synchronization of usual cameras as well as devices constructed by integrating optoelectronic components situated in order to provide the desired layout of viewpoints and then synchronized using specifically designed electronics. Lastly, these capture tools differ by the target use of the multiview media they capture (direct multiscopic visualization, FTV, reconstruction, refocus, etc.) which has a direct impact on the compromise between the number of views and their resolutions to maintain an acceptable volume of pixels to be captured, transmitted and stored.

These close multiview capture tools (either assembled or integrated) are often known as "camera arrays" (grids or linear layouts of cameras or viewpoints) and "plenoptic" systems or cameras. Camera arrays are generally focused on capturing multiple images with significant resolution for the depth reconstruction and 3D and/ or interactive visualization (FTV) whilst plenoptic systems generally aim to capture the "light field", and are more balanced in terms of the number of views and resolution to extract interpolated views (FTV) or variable focus images (refocus) as well as, sometimes, depth reconstructions. This classification is more nuanced than it seems because the similarity of their shooting geometries and improvements in capture and pixel processing capabilities volumetric capabilities tend to bring closer those ratios number of views/ number of pixels per view and therefore mean that intended applications are accessible by both types of system. This classification could, however, soon be historical artifact related to the appearance in successive waves of these technologies as well as their original objectives.

Undeniably, the first devices proposed fell within the class of linear viewpoints arrangements. Initially limited to capturing static scenes (in terms of composition as well as lighting), the very first systems achieved multiple perspective captures by controlling sequential positions of a still camera, as developed by Stanford University [LH96]. They were quickly overtaken by multisensor devices taking images of the same dynamic scene simultaneously, such as that proposed by Dayton Taylor in 1996 [Tay96], and/ or in low level and controlled desynchronization such as the system developed by Manex Entertainment for the film *The Matrix*. The majority of these devices were often designed and build specifically for their desired function: the MERL 3DTV project by Mitsubishi [MP04] positioned sixteen cameras on a rail to produce multiscopic content designed for their ad hoc autostereoscopic screens whilst the University of California in San Diego, with Mitsubishi [JMA06] used a rail with eight cameras for an automatic video matting application. Several prototypes of integrated devices have also been proposed for specific applications. We can, for example, cite the cameras with eight viewpoints developed in Reims, France [PCPD$^+$10], which are illustrated in figure 3.3, and which were specifically designed to produce multiscopic content with controlled distortion (see chapter 4) for autostereoscopic screens on the market.

These linear layouts have, in addition, also been extended by several laboratories to more complex systems of 2D grids of cameras. The most well known is probably that created by Stanford University[23] [WJV$^+$05] which has been used for multiple applications, notably aimed at FTV and refocus. It is composed of a variable number of cameras (usually more than a hundred) organized according to various configurations in planar or piecewise planar 2D grids. Another 2D grid, albeit irregular, has been developed by the Carnegie Mellon university [ZC04] with 48 cameras in individual horizontal and vertical positions controlled to optimize the calculation of depth in order to generate the desired perspective (FTV). We can also cite Sony in partnership

[23] http://graphics.stanford.edu/projects/array/.

with Columbia University [NZN07] who have proposed flexible and stretchable 1D and 2D grids, composed of elastic supports on which twenty cameras are fixed in regular positions (at rest state). The deformation of the support therefore modifies the system's configuration to adapt to the situation and the desired requirements (more or less panoramic mosaicing in [NZN07]).

The emergence of grids has also enabled research dealing with ray-space associated with plenoptic function, notably summarized by [AB91]. This plenoptic function (an aggregation of the Latin *plenus* – complete – and optics) is the function which gives the light intensity of all the rays in a scene. Yielding real value, it is defined for seven real variables; three for the position of a point of the ray, two for its 3D direction of propagation, one for the wavelength from which we measure intensity and the last for the point in time of this measure:

$$\begin{aligned} \mathcal{P} \quad \mathbb{R}^3 \times \mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}/\pi\mathbb{Z} \times \mathbb{R}^+ \times \mathbb{R} &\longmapsto \mathbb{R}^+ \\ ((x,y,z),(\phi,\theta),\lambda,t) &\longrightarrow \mathcal{P}(x,y,z,\phi,\theta,\lambda,t) \end{aligned} \tag{3.7}$$

Usually, this function is reduced to five variables by externalizing the wavelength in the result which becomes a spectrum and by considering the intensity to be constant at the time of measure along the whole length of the ray[24]. According to this hypothesis, all the points in the ray deliver almost the same spectrum at the time studied and we can therefore reduce this redundancy by suppressing one of the space variables. In practice, we commonly select coplanar points by no longer "managing" the rays parallel to this ray capturing plane. This gives:

$$\begin{aligned} \mathcal{P} \quad \mathbb{R}^2 \times \mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}/\pi\mathbb{Z} \times \mathbb{R} &\longmapsto \mathbb{R}^{+^{\mathbb{R}^+}} \\ ((x,y),(\phi,\theta),t) &\longrightarrow \mathcal{P}(x,y,\phi,\theta,t) \equiv \text{ spectrum } \mathcal{S}(\lambda) \end{aligned} \tag{3.8}$$

The domain's dimension can be again reduced to four by fixing the time of study or by transferring it in the result which becomes a temporal spectrum:

$$\begin{aligned} \mathcal{P} \quad \mathbb{R}^2 \times \mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}/\pi\mathbb{Z} &\longmapsto \mathbb{R}^{+^{\mathbb{R}^+ \times \mathbb{R}}} \\ ((x,y),(\phi,\theta)) &\longrightarrow \mathcal{P}(x,y,\phi,\theta) \equiv \text{ temporal spectrum } \mathcal{S}(\lambda,t) \end{aligned} \tag{3.9}$$

Digitalizing the reduced plenoptic function involves spatial, angular, spectral and temporal windowing and sampling operations followed by quantification of the intensities which limit the domain as well as the value space. These operations create a temporal series of 4D digital signals indexed by the indices i,j (connected to x,y) from the capture points arranged in a grid and the coordinates s,t of the image pixel captured (in i,j), representative of the direction ϕ,θ of the ray measured in i,j,s,t. For each sample they contain a set of intensities quantified for a discrete number of spectral bands (generally 3 - RGB). These light fields can be easily obtained from the data captured by a camera array by simply stacking up the views captured according to the grid's layout:

$$\mathcal{LF}\,[s,t,i,j] \equiv \mathit{Quantify}\,(\mathcal{P}(x(i,j),y(i,j),\phi(i,j,s,t),\theta(i,j,s,t))) \tag{3.10}$$

The growing attraction for this multiview capture representation and, specifically for its resulting models and applications (FTV, refocus, to name but a few), has lead to the arrival of dedicated optics such as that proposed by Todor Georgiev from Adobe-Qualcomm[25] and

[24] Given that we temporally sample time at a step dt and then that the light intensity if transported to the speed of light c yielding $\mathcal{I}(x,t) = \mathcal{I}(x0, t-(x-x0)/c)$, this hypothesis is reasonable if the maximum width of the scene is slightly inferior to the distance travelled by a photon between two time steps, namely 299 792 458.dt m $\approx$ 12 491 km at 24 Hz, 2 998 km at 1 kHz or even 300 m at 1 MHz[1].

[25] www.wired.com/gadgetlab/2007/10/adobe-shows-off/.

integrated solutions, such as the "plenoptic cameras" proposed in recent years by companies such as Raytrix[26] or Lytro[27] (see figure 3.3). These cameras generally include a microlens grid in front or behind the lens in order to separately capture, after deviation, the light rays which are combined in a standard camera (see figure 3.4 for an illustration with a lenticular array at the back wall of the darkroom). If the object is captured in the focus plane (example B in figure figure 3.4), instead of a clear pixel, we obtain a homogenous micro-image which is synonymous with the object's position being in the focus plane. Otherwise (examples A and B), we obtain, instead of a blurred pixel, a local sampling of the object which, coupled with those of the neighboring capture positions, allows reconstructing the points outside of the focus plane. Other approaches, notably that by Mitsubishi[VRA$^+$07] [28], replace the lenticular array with a printed mask similar to parallax barriers. As a result, the debate between masks and microlenses, well known with autostereoscopic displays, also applies to plenoptic cameras.

Figure 3.3: Examples of integrated cameras: left, a Cam-Box prototype camera with eight integrated perspectives developed by 3DTV Solutions and the University of Reims and, right, the Lytro plenoptic camera

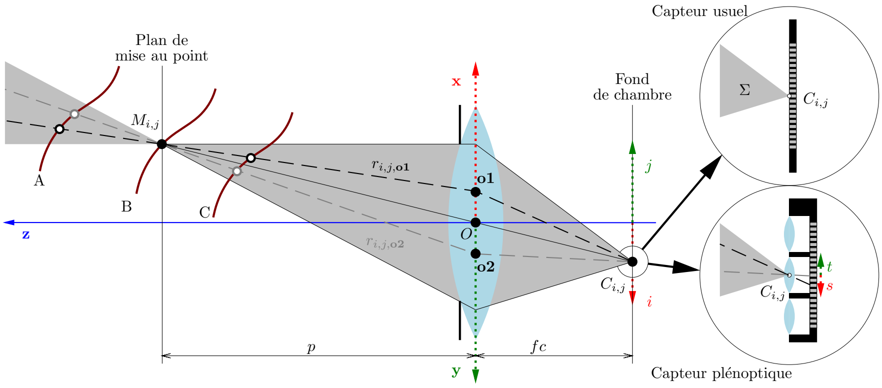


Figure 3.4: Differences between standard and plenoptic cameras:
from above (axes $\mathbf{x}, j, s$) or the side (axes $\mathbf{y}, j, t$)
the rays converging as a single point at the back wall of the darkroom are summed in the first and differentiated by refraction and sampling in the second

There has also been a recent tendency to miniaturize small grids within new integrated components, designed specifically for mobile terminals. The Californian company Pelican Imaging

[26] www.raytrix.de/index.php/Cameras.html.
[27] https://www.lytro.com/camera.
[28] http://web.media.mit.edu/~raskar//Mask/.

has produced a 5 × 5 microgrid component which is the size of a current monoview sensor[29].

3.3.2 Principal uses

Linear layouts of different viewpoints allow, by simple selection (or even interpolation) of a specific viewpoint, the effect of camera movement around a frozen or slow-motion scene. These technologies, known as bullet time, were largely brought to the fore in 1999 by the film *The Matrix*. It has since been used by a number of companies using more or less integrated proprietary systems which can be used with varied and occasionally surprising applications such as surfing, for example[30].

With the emergence of multiscopic visualization devices (see chapter 14), the question of creating adapted content using real capture has been developed, notably leading to several improvements in camera arrays. Linear layouts have also focused on autostereoscopic devices with a simple horizontal parallax. Similarly, grids have also been used for double parallax devices, known as "integral imaging displays" in reference to its precursor, "integral photography" proposed [Lip08b] and then experimentally validated [Lip08a] in 1908 by Gabriel Lippmann.

The generation of intermediary viewpoints (FTV, "image based rendering", IBR) also had a strong influence on the emergence of different camera arrays. This technology is somewhat of an extension of the frozen time virtual camera technique using camera position interpolation. Its implementation is, however, different and relies either on a depth reconstruction to project the available views on the virtual camera (see chapter 9) or on a planar section of the light field (with the real, coordinates i, j fixed), yielding a digital signal which samples the reduced plenoptic function according to equation (3.10).

The strong redundancy of close-together multiple perspective captures in a single scene can provide a depth reconstruction with increased reliability. As the quality of both depth maps (or disparity maps with parallel geometry capturing) and occlusion detection is essential in related applications (such as FTV and AR), a number of teams have studied the opportunity to use these strong redundancies which imply additional new challenges. Multiple solutions have been proposed, seeking coherence between multiple binocular matches or directly examining multi-ocular matches across all views. Regardless of the approach, managing occlusions, which is accessible in multi-ocular vision, is an opportunity which remains difficult to manage. Chapter 7 provides a more detailed description of this area.

Similarly, the availability of strongly redundant views allowing for a global matching process has been used (see chapter 19) to create high dynamic range (HDR) capturing devices by post-processing, views captured with moderate but varied dynamic from different viewpoints. The allocation of different dynamic ranges to viewpoints is obtained by neutral filters of different densities or by distinct exposure time settings.

To conclude, let us present an example of application of multiview capture, either by grids or plenoptic cameras, which is surprising since the notion of depth of field, a crucial aspect of photography, seemed definitely set at shooting. The numerous multi-view capture as well as ray-space modeling have given rise to a flurry of activity relating to a new opportunity with highly promising possibilities: the choice to refocus post capture. This includes, for example:

- the selection of the focus plane (by averaging pixels from several perspectives corresponding to the rays geometrically issued from the same points in this plane);

[29] www.pelicanimaging.com/index.htm.

[30] www.core77.com/blog/technology/rip_curl_time-slice_camera_array_collaboration_lets_you_perceive_surfing_as_never_before_20925.asp.

- the choice of aperture and therefore depth of field (by selecting the neighboring viewpoints from which the averaged pixels are taken);
- the possibility of selecting an "all-in-focus" infinite depth of field (by selecting non averaged pixels, which corresponds to a pinhole camera);
- removing the foreground from some images, to show the partially hidden background, if it is far enough away to be visible from several other viewpoints.

3.3.3 Related databases

Without attempting to provide an exhaustive list, there are a number of databases created using the devices discussed in this section. The University of California in San Diego and Mitsubishi[31] deliver some capture in a linear layout with 8-view videos and a series of 120 to 500 still images. The Light Fields library at the University of Stanford[32] is full of highly varied multiple scenes captured in high resolution, often from several hundred viewpoints, created by moving the camera on robotized arms or the Stanford grid. This information is available as raw or modified data with calibration information and the possibility of interacting online with their light field form by selecting a perspective and handling refocus (choice of shutter and focus plane). This library completes and surpasses its predecessor[33] which proposed less complex series, both in terms of the number of views as well as their resolution. A simpler example is also available on Todor Georgiev's site [34], which contains a number of plenoptic images with several tens of millions of rays. Lastly, the University of Heidelberg maintains a library [35] of several synthesized light fields, accompanied by genuine depth information, as well as real scene captures by the Raytrix plenoptic cameras using a 9 × 9 grid.

[1]on fait aussi l'hypothèse que le medium est transparent (pas de solide, de brume ou de fumée dans la zone où sont les caméras), et a un indice de réfraction constant dans cette même zone (sinon les rayons sont tordus) mais c'est un détail. en dehors de cette zone, ou plus particulièrement dans le volume vu par toutes les caméras, il peut se passer n'importe quoi, mais on ne peut espérer reconstruire des points de vue que dans la partie de l'espace où les hypothèses sont vérifiées

3.4 Global or omnidirectional multiview systems

3.4.1 Technical description

In this section we will examine multi camera systems with spaced out and approximately convergent layout in order to "cover" with enough redundancy a scene volume large enough to encompass evolving objects and/ or actors. The first systems of this kind have been used for bullet time or motion capture techniques. "Global systems" used for frozen time are generally composed of a rail forming a curve representing the desired trajectory for the virtual camera (i.e. closed or not, not always planar or circular, etc.) often hosting a significant number of cameras with a viewing direction set according to that desired for the virtual camera at this place, and with controlled synchronization depending on the desired effect (frozen time or more or less slow-motion). In motion capture (MoCap) using video markers, for the most part one

[31] http://graphics.ucsd.edu/datasets/lfarchive/lfs.shtml.

[32] http://lightfield.stanford.edu/.

[33] http://graphics.stanford.edu/software/lightpack/lifs.html.

[34] www.tgeorgiev.net/Gallery/.

[35] http://hci.iwr.uni-heidelberg.de/HCI/Research/LightField/lf_archive.php.

uses fewer synchronized infrared cameras freely positioned, and a geometric calibration obtained by moving a target object bearing fixed markers.

The fairly intensive use of these techniques by the film and video games industries (whose business-model make it profitable), has raised a marked interest in a more advanced technology using markerless multiview capture with more varied results: 3D video. Proposed in 1997 [KRN97, MTG97] and intensively studied and developed since then, [MNT12], it allows the reconstruction within an entire sequence of the geometry as well as the texture of the object or actor being filmed to create an animated digital avatar of sufficient quality that it can be reused by synthesizing the image from loosely restricted angles.

This requires a synchronized multiview capture system with numerous viewpoints distributed around the scene space, characterized as the intersection of camera fields of view (see left of figure 3.5). The compromise between the number of cameras (completion) and the gap between cameras (precision of reconstruction) has been suggested by [KRN97] to be between nine and sixteen for a circular and regular layout placed at mid-height of the scene space with converging axis at the circle center (see top left of figure 3.5 for an example with twelve cameras). More complete solutions have also been proposed to reconstruct the top of objects by adding cameras overlooking the scene from above and then selecting layouts sampling more envenly the directions of capture (several circles at different heights with aerial cameras[36], domes[37] [38], in more ad hoc studio or outside layouts [KGT$^+$12][39]) with the number of cameras fluctuating depending on the applicative context from a few units (University of Surrey[39], *Max Planck Institute* [dAST$^+$08] or the "GrImage" project[40]) to several hundreds (1000 for the "Virtualized reality" project[41]).

These complex systems must also have networking, storage and calculatory capabilities in order to manage generated video streams and very precise geometric and colorimetrics calibration technologies. Lastly, controlling lighting conditions and simplifying objects outlining facilitates image processing. This renders these systems complex, delicate and costly and explains their normal use in dedicated rooms known as "3D video studios".

The "bullet time" market is principally structured around service providers[42] which operate proprietary systems whilst MoCap concerns also several companies[43] who distribute off-the-shelf solutions. With regards to 3D video, the service has developed with specialized production companies with 3D studios[44] whilst the commercialization of these systems is just beginning[45].

3.4.2 Principal uses

In this section we will not discuss at length frozen time or MoCap technologies as their fairly specific capturing systems position them at the edge of the scope of this book. Hence, the main use of "global multiview systems" concerns 3D video, which have witnessed a boom both in research and production, as noted in [MNT12] which focuses entirely on this technique. 3D video relies on complex systems including a number of cameras synchronized, distributed and calibrated in terms of geometry and colorimetry within a video stream transfer network with significant storage and calculation capabilities.

[36] Recover3D, a project, 2012-2014, run by XD Productions, see far right and bottom of figure 3.5.

[37] www.cs.cmu.edu/ virtualized-reality/page_History.html.

[38] The 3D-COFORM FP7 project 2007-2013, www.vcc-3d.eu/multiview and www.3dcoform.eu, digitalizing heritage for small objects exhibiting complex light/ matter interactions.

[39] www.surrey.ac.uk/cvssp/research/3d_video/index.htm.

[40] www.inrialpes.fr/grimage/.

[41] www.cs.cmu.edu/ virtualized-reality/.

[42] Such as Reel EFX www.reelefx.com/ and Time Slice www.timeslicefilms.com/#1.

[43] Such as Vicon (www.vicon.com/), Animazoo (www.animazoo.com/) and Moven (www.moven.com/).

[44] For example, XD Productions (www.xdprod.com/) and 4D View Solutions (www.4dviews.com/).

[45] 4D View Solutions www.4dviews.com/ has also been marketing solutions for some time.

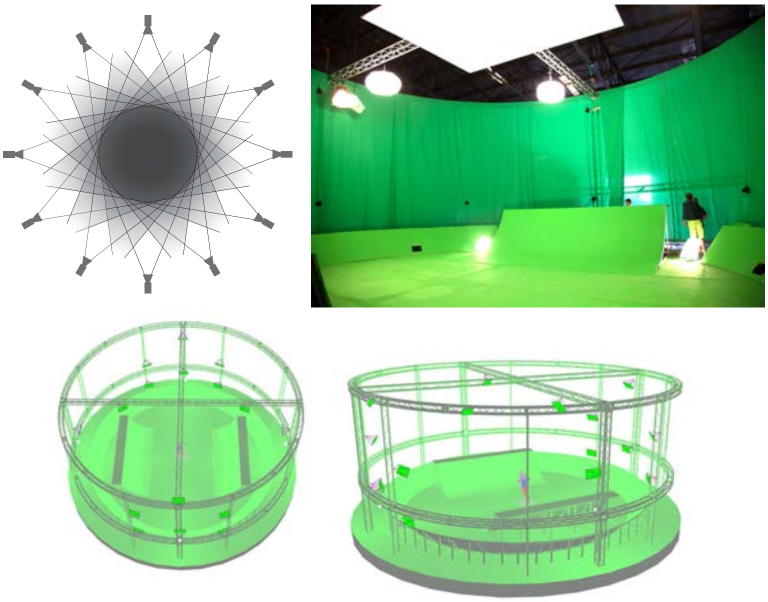


Figure 3.5: Examples of 3D video studios: from top left, circular arrangement of twelve cameras showing the scenic space used as an intersection of camera field depth zones (in light gray); top right and below, the studio of the Recover3d project[36]

The extraction of avatars' geometry from multiple video streams initially requires a precise geometric calibration of all cameras. This reconstruction can be operated according to three techniques classed as "model based" or, in contrast, free methods. The first class corresponds to searching the configuration of a predefined model which optimizes the geometric model's degrees of freedom so that its projections correspond to the images captured as closely as possible. The second contains two competing techniques; multiview stereo, which aims to reconstruct 3D points by triangulation using supposedly homologous pixels in different images and "silhouette based" methods which reconstruct the visual hull of the avatar by intersecting generalized cones supported by the outlines of its projections in all images. However, searching a predefined model configuration has shown a fairly fatal flaw in its construction; it lacks adaptability although it can, nevertheless, guide a silhouette-based reconstruction using fewer cameras ([dAST$^+$08], the "Free Viewpoint Video of Human Actors" project[46] [CTMS03]). Stereovision methods are sensitive to errors in colorimetric calibration and to specular reflections, are generally very costly in terms of computation time but can provide geometric information in concave zones where the visual hull is naturally convex. In contrast, visual hulls are easier to obtain, can be computed efficiently, are more reliable although these envelopes provide, by their very nature, only rough results in concave zones of the objects. The model-based techniques are often employed to digitize human actors. Among free methods (non model-based), even when applied to humans, "Visual Hull" techniques (examined in chapter 8) are often used in production due to their reliability, although their limitations have restricted their progression so far. It is for this reason that the complimentary combination of multiview stereo and silhouettes has inspired projects based on creating hybrids of them such as Recover3d[36]In which monoscopic and multiscopic cameras are distributed around the scene space to produce a robust geometric model (by integrating it

[46] www.mpi-inf.mpg.de/ theobalt/FreeViewpointVideo/.

into the visual hull) which is more detailed (through multiview stero reconstruction), notably in concave areas.

Once the geometric model has been reconstructed at each time step, it has to be given a temporally coherent visual content (texture) taken from the captured images. One may apply geometric models temporal tracking solutions (see chapter 8) to create semantic coherence between texture hooks, followed by video texturing techniques which involve locally mixing photometric information re-projected onto the geometric model from the images where this local zone is not hidden. Difficulties here relate to what decision to make when there are gaps between retro-projected data. These gaps can originate in geometric reconstruction faults, colorimetric calibration faults, as well as characteristics related to the scene itself such as reflections, or other specular phenomena. These complex visual phenomena are the basis of further study, such as the Light Stages series[47], which examines systems dedicated to capturing complex optical properties in a camera array context with lightning conditions modulation or, more recently, the 3D-COFORM project[38] which focuses on the high quality digitalization of heritage and cultural objects through capturing static objects in multiple lighting conditions (151 sources) from 151 viewpoints and different exposures to create HDR views (one per source/ viewpoint pair), thereby enabling mapping of optical properties in the form of bidirectional function textures (BFTs).

Video3D capture is more costly than MoCap because it is more complex. However, its results are far more versatile. Indeed, the producer and his/ her graphics technicians can, in post production, easily select the angles of view with few spatial limitations whilst editing the animated avatars acquired in these scenes (spatio-temporal movement/ deformation, duplication, transposition into other scenes, relighting[48].). These possibilities make these acquired avatars more re-usable and profitable, thereby reducing production costs. This creates a kind of technology that is both open to creativity and cheaper and is more accessible for televisual production. As a result, the digitalization of animated avatars is also of interest for other applicative domains such as culture[38], sport [KGT$^+$12] and collaborative telepresence [PDB$^+$10].

Lastly, a recent tendency, outside of the scope of this chapter, extrapolates the 3D video capabilities described previously, by targeting 3D reconstruction using non calibrated collective sources (such as web found amateur captures) in the form of photos [GSC$^+$07, Sna09] or videos ([BBPP10],the “Virtual Video Camera” project[49]).

3.4.3 Related databases

Several academic sites offer multiview sequences captured by their systems. The University of Surrey gives 8-view captures in a circular layout (www.ee.surrey.ac.uk/cvssp/visualmedia/visual-contentproduction/projects/surfcap), MIT proposes a number of complete data sets (images, exposure, results, etc.) which have been captured and processed according to [VBMP08] (http://people.-csail.mit.edu/drdaniel/mesh_animation/) and Inria Grenoble-Rhône-Alpes has made public its “4D repository” of several tens of data sets captured by their GrImage[1] system (http://4drepository.inrialpes.fr/).

3.5 Conclusion

This chapter has shown that multiview capture entails the use of varied and highly complex technologies. These technologies have opened up new perspectives on more creative post production processes which could revolutionize audiovisual production whilst offering further potential for

[47] http://gl.ict.usc.edu/LightStages/.

[48] A number of illustrations of this can be found on the XD Productions website www.xdprod.com/Xd Productions_RD.swf.

[49] http://graphics.tu-bs.de/projects/vvc/.

qualitative editing of recorded media post filming. They also provide an increasingly rich means of digitalizing our environment, as well as a number of other applicative fields requiring 3D reconstruction and/ or motion recognition. Whilst these technologies are currently mainly being developed as laboratory prototypes, as ad hoc systems for service providers or batch production devices, the importance of these applications will enable their commercial development, as shown by the arrival of plenoptic cameras and microgrids [1]for mobile devices.

Bibliography

[AB91] Edward H. Adelson and James R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20. MIT Press, Cambridge, MA, Etats-Unis, 1991.

[BBPP10] Luca Ballan, Gabriel J. Brostow, Jens Puwein, and Marc Pollefeys. Unstructured video-based rendering: interactive exploration of casually captured videos. In *Proceedings ACM SIGGRAPH*, pages 87:1–87:11, Los Angeles, CA, Etats-Unis, July 2010.

[CTMS03] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. In *Proceedings ACM SIGGRAPH*, pages 569–577, San Diego, CA, Etats-Unis, July 2003.

[dAST$^+$08] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *Proceedings ACM SIGGRAPH*, volume 27, pages 98:1–98:10, Los Angeles, CA, Etats-Unis, August 2008.

[DB10] Frédéric Devernay and Paul Beardsley. Stereoscopic cinema. In Rémi Ronfard and Gabriel Taubin, editors, *Image and Geometry Processing for 3-D Cinematography*, volume 5 of *Geometry and Computing*, chapter 2, pages 11–51. Springer, Heidelberg, Allemagne, 2010.

[ENO05] M. Emoto, T. Niida, and F. Okano. Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television. *Journal of Display Technology*, 1(2):328–340, December 2005.

[GSC$^+$07] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz. Multi-view stereo for community photo collections. In *Proceedings ICCV, IEEE International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Bresil, October 2007.

[JMA06] N. Joshi, W. Matusik, and S. Avidan. Natural video matting using camera arrays. In *Proceedings ACM SIGGRAPH*, volume 25, pages 779–786, Boston, MA, Etats-Unis, July 2006.

[KGT$^+$12] Hansung Kim, J.-Y. Guillemaut, T. Takai, M. Sarim, and A. Hilton. Outdoor dynamic 3-d scene reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(11):1611–1622, November 2012.

[KRN97] Takeo Kanade, Peter Rander, and P. J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE MultiMedia*, 4(1):34–47, January 1997.

[LH96] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings ACM SIGGRAPH*, pages 31–42, Nouvelle Orléans, LA, Etats-Unis, August 1996.

[Lip08a] M. G. Lippmann. Epreuves réversibles donnant la sensation du relief. *Journal de Physique Théorique et Appliquée*, 7(1):821–825, November 1908.

[Lip08b] M. G. Lippmann. Epreuves réversibles. photographies intégrales. *Comptes Rendus de l'Académie des Sciences*, 146(9):446–451, March 1908.

[Lip82] Lenny Lipton. *Foundations of the Stereoscopic Cinema*. Van Nostrand Reinhold, New York, NY, Etats-Unis, 1982.

[Men09] Bernard Mendiburu. *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Focal Press, Burlington, MA, Etats-Unis, 2009.

[Men11] Bernard Mendiburu. *3D TV and 3D Cinema: Tools and Processes for Creative Stereoscopy*. Focal Press, Waltham, MA, Etats-Unis, 1^{e} edition, 2011.

[MMN99] Susana Marcos, Esther Moreno, and Rafael Navarro. The depth-of-field of the human eye from objective and subjective measurements. *Vision Research*, 39(12):2039–2049, June 1999.

[MNT12] T. Matsuyama, S. Nobuhara, and T. Takai. *3D Video and Its Applications*. SpringerLink : Bücher. Springer, Londres, Royaume-Uni, 2012.

[MP04] W. Matusik and H. Pfister. 3d tv: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. In *Proceedings ACM SIGGRAPH*, volume 24, pages 814–824, Los Angeles, CA, Etats-Unis, August 2004.

[MTG97] Saied Moezzi, Li-Cheng Tai, and Philippe Gerard. Virtual view generation for 3d digital video. *IEEE MultiMedia*, 4(1):18–26, January 1997.

[NZN07] Y. Nomura, L. Zhang, and S.K. Nayar. Scene Collages and Flexible Camera Arrays. In *Proceedings EGSR, Eurographics Symposium on Rendering*, June 2007.

[PCPD^{+}10] Jessica Prevoteau, Sylvia Chalençon-Piotin, Didier Debons, Laurent Lucas, and Yannick Remion. Multi-view shooting geometry for multiscopic rendering with controlled distortion. *International Journal of Digital Multimedia Broadcasting (IJDMB), special issue Advances in 3DTV: Theory and Practice*, 2010:1–11, March 2010.

[PDB^{+}10] Benjamin Petit, Thomas Dupeux, Benoit Bossavit, Joeffrey Legaux, Bruno Raffin, Emmanuel Melin, Jean-Sébastien Franco, Ingo Assenmacher, and Edmond Boyer. A 3d data intensive tele-immersive grid. In *Proceedings MM, international conference on Multimedia*, pages 1315–1318, Florence, Italie, 2010. ACM, New York, NY, Etats-Unis.

[Sna09] Keith N. Snavely. *Scene reconstruction and visualization from internet photo collections*. PhD thesis, Seattle, WA, Etats-Unis, 2009.

[Tay96] D. Taylor. Virtual camera movement: The way of the future ? *American Cinematographer*, 77(9):93–100, 1996.

[UH07] Kazuhiko Ukai and Peter A. Howarth. Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations. *Displays*, 29(2):106–116, March 2007.

[VBMP08] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *Proceedings ACM SIGGRAPH*, volume 27, pages 97:1–97:9, Los Angeles, CA, Etats-Unis, August 2008.

[VRA$^+$07] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack. Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. In *Proceedings ACM SIGGRAPH*, volume 26, San Diego, CA, Etats-Unis, July 2007.

[WJV$^+$05] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. In *Proceedings ACM SIGGRAPH*, pages 765–776, Los Angeles, CA, Etats-Unis, July 2005.

[YEM04] Sumio Yano, Masaki Emoto, and Tetsuo Mitsuhashi. Two factors in visual fatigue caused by stereoscopic HDTV images. *Displays*, pages 141–150, November 2004.

[ZC04] Cha Zhang and Tsuhan Chen. A self-reconfigurable camera array. In *Proceedings EGSR, Eurographics Workshop on Rendering Techniques*, pages 243–254, Norköping, Suéde, June 2004. Eurographics Association, Aire-la-Ville, Suisse.