Systèmes d'acquisition multivues

Frédéric Devernay, Yves Pupulin, Yannick Rémion

Chapitre 3

Systèmes d'acquisition multivues

3.1. Introduction : qu'est-ce qu'un système d'acquisition multivues ?

L'acquisition multivues, objet de ce chapitre, concerne la capture de données vidéo synchronisées représentant différents points de vue d'une même scène. *A contrario* des systèmes de vidéosurveillance qui déploient de multiples caméras pour couvrir visuellement avec peu de redondance un espace à surveiller de grande étendue, les matériels, dispositifs ou systèmes ici visés couvrent, depuis plusieurs points de vues, un même espace physique, souvent assez réduit, pour profiter de la redondance des images à des fins :

– de visualisation en relief stéroscopique ou multiscopique des vidéos captées ;

– de reconstruction/virtualisation de scène réelle :
  - reconstruction 2,5D de carte de profondeur depuis un point de vue donné,
  - reconstruction 3D texturée de modèles numériques, avatars des objets réels,
  - capture de mouvement pour animation réaliste d'acteurs virtuels ;

– d'ajustements divers et complémentaires en régie ou en postproduction :
  - « mosaïquage » de vues donnant une vue panoramique ou haute résolution,
  - caméra virtuelle se déplaçant à temps figé ou très ralenti (*bullet time*),
  - mixage réel/virtuel (« réalité augmentée » – RA),
  - interpolation de points de vue (*free viewpoint TV* – FTV),
  - modification *a posteriori* de la mise au point (*refocus*),
  - extension *high dynamic range* – HDR – de la dynamique des vidéos,
  - etc.

Chapitre rédigé par Frédéric DEVERNAY, Yves PUPULIN et Yannick REMION.

Selon les applications visées, le nombre, la disposition et le réglage des caméras peuvent grandement fluctuer. Les configurations les plus communes à ce jour sont :

– les « systèmes binoculaires » proposant deux points de vue proches ; ces systèmes permettent la visualisation en relief stéréoscopique (généralement à lunettes) et la reconstruction de profondeur avec les postproductions associées (RA, FTV) ;

– les « systèmes multivues latéraux ou directionnels[1] » disposant de multiples points de vue proches (généralement régulièrement espacés), tous placés d'un même côté de la scène ; ces systèmes permettent une visualisation relief autostéréoscopique, des effets limités de « temps figé » et une reconstruction de profondeur ou reconstruction 3D « directionnelle » plus robuste que pour le cas binoculaire avec les mêmes postproductions accessibles (RA, FTV). La multiplication des points de vue permet aussi la différenciation des réglages de chaque caméra qui, avec la forte redondance des captures, rend d'autres posttraitements accessibles (*refocus* et HDR notamment) ;

– les « systèmes multivues englobants ou omnidirectionnels[1] » disposant leurs multiples points de vue tout autour de l'espace cible ; ces systèmes sont principalement destinés au *bullet time* à large secteur angulaire, à la reconstruction 3D et à la capture de mouvement (*motion capture* – MoCap).

A côté de ces solutions purement vidéo, il est utile de mentionner les systèmes hybrides qui ajoutent un capteur de profondeur (« Z-cam ») au(x) capteur(s) vidéo. La profondeur capturée permet théoriquement un accès direct à la plupart des postproductions envisagées. Le nombre de capteurs vidéo comme la résolution de la capture de profondeur et ses limitations spatiales peuvent cependant limiter certains de ces post-traitements. Ces systèmes hybrides ne sont pas détaillés dans cet ouvrage.

Tous ces matériels partagent le besoin de synchronisation et de calibration (parfois même de rectifications géométriques et/ou colorimétriques) des captures par les différentes caméras ou Z-cam et disposent souvent de capacités annexes concernant :

– l'enregistrement des signaux de tous les capteurs sans perte de données ;

– le traitement de l'ensemble des données en temps réel, ce qui exige une infrastructure de calcul importante (avec en général distribution des calculs).

Ce chapitre va présenter les principales configurations de capture multivues purement vidéo ci-dessus citées en se basant sur des exemples significatifs de réalisations et en indiquant les usages qui en sont faits. A chaque fois, seront proposées des bases de données accessibles permettant au lecteur d'accéder à des médias acquis par des dispositifs relevant de la catégorie étudiée.

1. Appellation proposée pour cet ouvrage.

3.2. Systèmes binoculaires

3.2.1. *Description technique*

La prise de vue vidéo binoculaire, aussi appelée stéréoscopique ou plus récemment « 3D stéréoscopique » (3D-S), nécessite l'utilisation de deux caméras[2], reliées par un dispositif mécanique rigide ou articulé appelé « rig stéréoscopique ». Les images captées peuvent être soit destinées à être projetées telles quelles sur un dispositif d'affichage stéréoscopique (écran de cinéma ou de télévision 3D, le plus souvent) [DEV 10], soit utilisées pour extraire la géométrie 3D de la scène, sous la forme d'une carte de profondeur, par des méthodes de vision stéréoscopique par ordinateur.

3.2.1.1. *Géométrie de prise de vue*

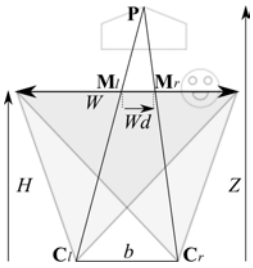
La prise de vue est effectuée avec deux caméras de mêmes paramètres optiques (focale, distance de mise au point, temps d'exposition, etc.), et visant approximativement dans une même direction orthogonale à la droite reliant leurs centres optiques, appelée ligne de base ou *baseline*. Les axes optiques peuvent être parallèles ou convergents.

Idéalement, pour simplifier la mise en correspondance stéréoscopique, les deux axes optiques devraient être strictement parallèles, orthogonaux à la ligne de base, et les plans images des deux caméras devraient être confondus. Dans cette situation, les points des images qui se correspondent se trouvent alors à la même ordonnée dans les deux images. Cependant, si les caméras sont convergentes (c'est-à-dire que les axes optiques convergent à une distance finie), ou si l'alignement est approximatif, les images issues des caméras peuvent être rectifiées (voir section 5.4) de manière à se ramener à cette situation idéale. La rectification est d'ailleurs une des phases de post-production de films stéréoscopiques (voir section 3.2.2.1).

Les principaux paramètres des dispositifs de capture comme de visualisation stéréoscopiques sont décrits figure 3.1. Soient b, W et H les paramètres de la caméra stéréoscopique, et Z la distance d'un point 3D au plan passant par la ligne de base stéréoscopique et parallèle aux plans images. Les triangles $\mathbf{M_l P M_r}$ et $\mathbf{C_l P C_r}$ sont homothétiques, en conséquence : $(Z - H)/Z = dW/b$. Cela permet d'exprimer simplement les relations entre la disparité stéréoscopique d, exprimée en fraction de la largeur W de l'image, et la distance Z de manière comparable à celle du chapitre 7 :

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

2. En photographie, lorsque la scène est fixe, on peut se satisfaire d'un appareil que l'on fait coulisser sur une réglette entre les prises de vues droite et gauche.



Symbole	Caméra	Affichage
$\mathbf{C}_l$, $\mathbf{C}_r$	centre optique	centre optique (yeux)
$\mathbf{P}$	point de la scène	point 3D perçu
$\mathbf{M}_l$, $\mathbf{M}_r$	points images de $\mathbf{P}$	points de l'écran
b	interoculaire	interoculaire (yeux)
H	distance de convergence	distance de l'écran
W	largeur du plan de convergence	taille de l'écran
Z	profondeur réelle	profondeur perçue
d	disparité gauche-droite (en fraction de W)	

Figure 3.1. *Géométrie du dispositif de capture stéréoscopique et celle du dispositif d'affichage stéréoscopique peuvent être décrites par le même faible nombre de paramètres*

3.2.1.2. *Distorsions géométriques perçues*

Si la vidéo stéréoscopique est destinée à être projetée sur un dispositif d'affichage stéréoscopique dont les paramètres sont b', W' et H', la profondeur Z' perçue par stéréoscopie[3] peut se calculer en fonction de la disparité d (équation (3.2)). En éliminant la disparité d de (3.1) et (3.2), on obtient en (3.3) la relation entre la profondeur réelle Z et la profondeur perçue Z' qui sera étendue au cas multiscopique au chapitre 4 :

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

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

Il y a divergence oculaire lorsque $Z' < 0$ ($d' > \frac{b'}{W'}$), soit pour une disparité absolue sur l'écran plus grande que l'interoculaire du spectateur). En général, les objets réels à l'infini ($Z \to +\infty$) sont perçus à distance finie ou causent de la divergence selon que $\frac{W'}{b'}\frac{b}{W}$ est plus petit ou plus grand que 1. On considère qu'une divergence oculaire de l'ordre de $0,5°$ est acceptable sur de courtes durées, et cela est utilisé par les stéréographes pour augmenter artificiellement l'espace disponible en profondeur derrière l'écran de cinéma.

Dans le cas de la télévision 3D, les limites de disparité dues au conflit entre vergence et accommodation [EMO 05, UKA 07, YAN 04] rendent très inconfortables les

3. La stéréoscopie se combine avec de nombreux autres indices monoculaires pour aboutir à la perception 3D de la scène [LIP 82] : lumière et ombre, taille relative, interposition, gradient de texture, perspective aérienne, perspective, flou, etc.

disparités importantes (positives ou négatives). La profondeur de champ de l'œil humain est de l'ordre de 0,3 δ (dioptries) dans les situations normales[4], ce qui donne pour une mise au point sur un écran placé à 3 m, une profondeur de champ allant de $1/(\frac{1}{3}+0,3) \approx 1,6$ m à $1/(\frac{1}{3}-0,3) = 30$ m. En pratique, les règles de production sont beaucoup plus restrictives : les programmes de TV 3D sont produits avec des disparités allant de -1 % à $+2$ % de la largeur de l'écran[5] pour rester dans cette zone de confort[6], avec temporairement des disparités allant de $-2,5$ % à $+4$ %, ce qui exclut totalement d'atteindre la limite de divergence pour des dispositifs de projection privés.

On remarque également que, pour que la profondeur perçue soit identique à la profondeur réelle ($Z' = Z$), il faut que tous les paramètres soient égaux, et on parle alors de configuration « orthostéréoscopique » (cette configuration est souvent utilisée pour les films IMAX 3D, puisque le dispositif de projection est de géométrie connue). Pour un plan 3D fronto-parallèle placé à une distance Z, on peut calculer le facteur d'échelle s entre les distances dans ce plan et les distances dans le plan de convergence : $s = H/Z$. On peut également calculer le rapport d'échelle image σ' qui explique combien un objet placé à la profondeur Z ou à la disparité d a l'air d'être agrandi ($\sigma' > 1$) ou réduit ($\sigma' < 1$) dans les directions X et Y par rapport aux objets se situant dans le plan de convergence ($Z = H$) :

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

Bien sûr, pour les objets dans le plan de l'écran ($d = 0$), on a $\sigma' = 1$. On note également que la relation entre Z et Z' n'est linéaire que si $W/b = W'/b'$, auquel cas $\sigma' = 1$ et $Z' = ZH'/H$. On parle alors de configuration « orthoplastique » (une configuration orthostéréoscopique est *a fortiori* orthoplastique).

Un petit objet de largeur ∂X et profondeur ∂Z, placé en Z, est perçu comme un objet de dimensions $\partial X' \times \partial Z'$ à la profondeur Z', et la rondeur (*roundness factor*) ρ mesure combien les proportions de l'objet sont modifiées :

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

Dans le plan de l'écran ($Z = H$ et $Z' = H'$), la rondeur se simplifie à :

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

---

4. Des études plus précises [MAR 99] montrent qu'elle dépend également de paramètres comme le diamètre pupillaire, la longueur d'onde et la composition spectrale.

5. Les disparités négatives correspondent à des points plus proches que l'écran, les disparités positives à des points plus éloignés.

6. Voir par exemple les règles de production de la chaîne anglaise Sky 3D : www.sky.com/shop/tv/3d/producing3d.

Une rondeur égale à 1 signifie qu'une sphère est perçue exactement comme une sphère, une rondeur plus petite qu'elle est perçue comme un disque aplati dans le sens de la profondeur, et une rondeur plus grande qu'elle est perçue comme un ellipsoïde allongé dans le sens de la profondeur. La rondeur d'un objet dans le plan de l'écran est égale à 1 si et seulement si $b'/b = H'/H$, et, pour qu'elle le soit partout dans l'espace, il faut que $b'/b = W'/W = H'/H$. Ainsi, les seules configurations géométriques qui conservent la rondeur partout sont identiques à la configuration d'affichage à un facteur d'échelle : ce sont les configurations « orthoplastiques ». Même si la géométrie du dispositif d'affichage est connue lors du tournage, cela impose des contraintes très fortes sur la manière de tourner un film, qui peuvent être très difficiles à suivre dans beaucoup de situations (par exemple lors du tournage d'événements sportifs ou de documentaires animaliers). D'autre part, comme l'interoculaire du spectateur b' est fixé, cela signifie qu'un film ne pourrait être projeté que sur un écran d'une taille donnée W' placé à une distance donnée H', ce qui est en contradiction avec la grande variabilité des dispositifs de projection et des salles de cinéma. On parle parfois de configurations « hyperplastiques » ou « hypoplastiques » lorsque la rondeur est respectivement plus grande ou plus petite que 1. On remarque également que la rondeur dans le plan de l'écran augmente lorsque l'on s'éloigne de l'écran, et qu'elle est indépendante de la taille de l'écran, ce qui est contre-intuitif : la plupart des spectateurs imaginent percevoir « plus de 3D » en approchant d'un grand écran.

Une autre remarque importante est qu'un film, tourné pour avoir une certaine rondeur pour un écran de cinéma situé en moyenne à 15 m, verra sa rondeur divisée par cinq une fois projeté sur un écran de TV 3D placé à 3 m, ce qui explique en partie l'insatisfaction des spectateurs de TV 3D. Cet effet peut être contre-balancé par une postproduction spécifique des médias destinés à la projection privée (*home cinema*), par exemple pour la version Blu-ray 3D, mais rares sont les titres qui bénéficient d'un tel traitement. Bien entendu, cette réduction de la rondeur est en partie compensée par les indices de profondeur monoscopiques. D'ailleurs, la rondeur utilisée dans les films de cinéma 3D est en réalité comprise le plus souvent entre $0,3$ et $0,6$, selon l'effet dramatique désiré [MEN 09], afin de favoriser le confort visuel du spectateur.

3.2.2. *Usages principaux*

3.2.2.1. *Cinéma et télévision 3D*

Les *rigs* de cinéma et de télévision sont majoritairement des systèmes lourds, et utilisent souvent un miroir semi-réfléchissant pour atteindre des entraxes de caméras plus courts que le diamètre des objectifs [MEN 11] (voir gauche de la figure 3.2). Quelques fabricants proposent aujourd'hui des caméras stéréoscopiques intégrées légères semi-professionnelles, mais leur champ d'utilisation est réduit, notamment à cause du fait que l'entraxe de ces caméras est généralement fixe, alors que la mise en

scène stéréoscopique nécessite un réglage adéquat de tous les paramètres de la stéréoscopie : ajouter une deuxième caméra à côté de la première ne suffit pas pour tourner en 3D-S.

3.2.2.1.1. La stéréoscopie, nouvel art, différent de la cinématographie

Le cinéma en 2D, pour exister, a dû (i) étudier le fonctionnement du cerveau afin de le tromper en lui faisant croire qu'une succession d'images fixes était en réalité un mouvement, (ii) répertorier, à partir de l'expérience acquise en photographie, les techniques permettant de réaliser cette illusion et élaborer une chaîne cinématographique complète, et (iii) inventer les paramètres d'un nouvel art, ce qui est l'affaire des artistes impliqués dans la fabrication des films, secondés par des ingénieurs fabriquant les outils correspondant aux nouvelles pratiques artistiques.

La stéréoscopie est à la fois en continuité et en rupture avec la cinématographie car, comme cette dernière vis-à-vis de la photographie, elle doit s'appuyer sur les techniques existantes et en développer d'autres. Pour ce faire, il est indispensable de :

– repartir de l'étude du cerveau et d'étudier comment le tromper, non temporellement mais, cette fois, spatialement, en recréant l'illusion d'un espace en trois dimensions alors qu'il ne s'agit en réalité que de deux images en 2D ;

– modifier pour la stéréoscopie les bons outils de prise de vue et de postproduction de la chaîne cinématographique et en fabriquer de nouveaux à partir de l'observation du fonctionnement cérébral afin d'assurer que cette nouvelle illusion soit confortable ;

– permettre l'invention d'une mise en scène basée sur les différents paramètres qui contribuent à créer cette illusion.

On considère connus les paramètres cinématographiques sur lesquels on peut jouer pour une prise de vue traditionnelle. Il reste à concevoir un outil à partir duquel les paramètres proprement stéréoscopiques seront utilisés pour cette nouvelle illusion. Basés sur le fonctionnement oculaire humain, ils devront permettre de simuler (i) la vergence couplée en général, dans l'observation humaine, à l'accommodation et (ii) la vision du relief résultant de l'entraxe entre les deux yeux, paramètre qui varie peu au long de l'existence de chaque individu et entre les individus.

Cependant, la simple adaptation de ce paramètre d'entraxe moyenné sur un échantillon de population ne peut-être, contrairement aux études ophtalmologiques, considérée comme suffisante. En effet, la stéréoscopie va jouer sur ces deux paramètres pour créer de l'émotion et du sens, exactement comme les objectifs utilisés sur une caméra ne cherchent pas à reproduire la vision perspective humaine mais à la déformer en fonction des choix de réalisation. Si l'on pousse aux extrêmes ces variations d'entraxe, on aura d'un côté la valeur 0 qui correspond à deux images en 2D identiques et à l'opposé, des distances interaxiales sans aucun rapport avec la physiologie. La NASA a par exemple réalisé des prises de vue stéréoscopiques de la terre avec près de soixante-dix mètres de distance entre les deux points de vue.

Pour la réalisation d'un *rig*, l'entraxe peut donc évoluer de 0 à la plus grande valeur utile pour certains types de prise de vue. En général, pour une configuration classique avec des comédiens, une variation de quelques millimètres à quelques centimètres répond à 90 % des besoins d'une fiction. Par conséquent, les *rigs* utilisés pour la cinématographie de personnages proches évolue de 0 à 100 mm au maximum. Enfin, pour les prises à grande distance, sur des nuages par exemple, la distance entre les deux caméras peut atteindre plusieurs mètres et le *rig* « côte-à-côte » utilisé sera fréquemment adapté aux besoins spécifiques.

3.2.2.1.2. Production assistée par ordinateur

Si les règles de re-création d'un univers en volume sont connues depuis le XIX^e^ siècle, la possibilité de maîtriser la mise en scène stéréoscopique en utilisant un *rig* est beaucoup plus récente et implique l'usage d'un ordinateur pour analyser les flux vidéo, et éventuellement en corriger les défauts. De la même façon, sachant qu'aucune technologie mécanique, optique, électronique n'est parfaite, il faudra impérativement corriger les images tournées le plus précisément possible avec un correcteur de relief opérant en temps réel pour la télévision et en post-production pour le cinéma, ce qui n'est envisageable que depuis l'invention des images numériques que l'on peut corriger pixel par pixel.

3.2.2.1.3. Un *rig* robotisé

Un *rig* doit utiliser des caméras synchronisées et des optiques aux mouvements de zoom, point et diaphragme parfaitement calibrés et synchronisés. Le *rig* lui-même est robotisé et comporte des motorisations permettant de régler l'entraxe et la vergence en temps réel, ainsi qu'un réglage d'assiettes visant à faire converger les deux axes optiques (les axes optiques doivent être coplanaires). Il arrive qu'un *rig* comporte plus de deux caméras comme ce fut le cas pour le film « La France entre ciel et mer » qui a été tourné par Binocle avec quatre caméras sur un hélicoptère (voir figure 3.2). Dans ce cas, l'appareillement des quatre zooms et des assiettes des quatre caméras demande une très grande expertise, puisque tous les centres optiques doivent être alignés le mieux possible. Des exemples de matériels permettant de piloter le *rig*, de contrôler la qualité en direct, et éventuellement de corriger les défauts géométriques et photométriques sont : TaggerLive et TaggerMovie de Binocle[7], STAN – *Stereoscopic Analyzer* – de Fraunhofer HHI, SIP – *Stereoscopic Image Processor* – de 3ality Technica[8], le processeur de correction temps réel MPES-3D01 – souvent appelé « 3DBox » – de Sony, et Pure de Stereolabs[9].

---

7. www.binocle.com.

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

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

Figure 3.2. *Exemples de* rigs *: à gauche, Brigger III de Binocle en configuration de studio,* rig *robotisé pour la TV 3D ; à droite,* rig *héliporté à quatre caméras utilisé par Binocle pour le film « La France entre ciel et mer »*

3.2.2.1.4. Postproduction stéréoscopique

Les outils de postproduction se sont également adaptés au cinéma 3D, et des algorithmes spécifiques à la stéréoscopie ont été intégrés dans ces logiciels : rectification, interpolation de points de vue et changement d'entraxe, conversion 2D vers 3D, égalisation colorimétrique des deux flux, production d'une carte de profondeur pour la composition avec des scènes 3D, etc. Parmi ces outils, on peut citer Nuke, et en particulier les *plug-ins* Ocula (*the Foundry*)[10], DisparityKiller (Binocle), et Mistika Post (SGO)[11].

3.2.2.2. *Reconstruction de profondeur*

Les systèmes binoculaires destinés à produire par reconstruction stéréoscopique des données 3D « partielles »[12] sont en général beaucoup moins complexes que ceux utilisés pour le cinéma ou la télévision. Ce sont le plus souvent des systèmes légers, peu encombrants, et consommant peu d'énergie, afin d'être utilisables par exemple sur un véhicule ou un robot mobile, et ils sont presque toujours à entraxe et à focale fixes, afin de simplifier leur calibration.

La plupart de ces systèmes utilisent des caméras monochromes, l'information de luminance seule étant en général suffisante pour la mise en correspondance stéréoscopique, mais la couleur peut apporter des fonctionnalités supplémentaires, comme la possibilité d'utiliser la couleur pour des tâches de segmentation (par exemple de la couleur de peau) ou de reconnaissance d'objets. Les caméras utilisées sur ce genre de système étant en général monocapteur, l'utilisation de la couleur implique une diminution de la résolution spatiale des images et donc de la précision de la profondeur reconstruite.

---

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

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

12. Au sens où elles forment depuis le point de vue utilisé un front de profondeur non assurément continu et assurément ouvert.

Le choix de la valeur d'entraxe optimale pour la reconstruction est un sujet controversé, mais des règles simples permettent de prévoir la précision finale. La précision de la disparité d obtenue par l'algorithme de mise en correspondance stéréoscopique peut être supposée constante sur l'image (disons $0,5$ pixels). L'erreur sur la profondeur reconstruite Z s'obtient en dérivant l'équation (3.1) : $\partial Z/\partial d = bHW/(b - dW)^2$, soit $\partial Z/\partial d = Z^2W/(bH)$. L'erreur croît donc comme le carré de la distance, et diminue théoriquement avec l'entraxe b, ce qui pousserait à choisir un entraxe le plus grand possible. Cependant, lorsque l'on augmente l'entraxe, la mise en correspondance stéréoscopique est plus difficile, et la précision de la disparité d se dégrade fortement lorsque le rapport b/H augmente. L'expérience montre qu'en règle générale un rapport b/H situé entre $0,1$ et $0,3$ donne un compromis raisonnable entre précision de la mise en correspondance stéréoscopique et la précision de la reconstruction de profondeur.

On peut utiliser n'importe quelle paire de caméras rigidement liées et synchronisées[13] pour faire de la reconstruction de profondeur par stéréoscopie (la bibliothèque logicielle OpenCV fournit des fonctions de calibration, de mise en correspondance stéréoscopique et de reconstruction 3D simples à utiliser).

Des systèmes commerciaux « clé en main » sont également disponibles : ils ont l'avantage d'être solidement construits, précalibrés ou facilement calibrables, et de proposer des algorithmes de mise en correspondance stéréoscopiques optimisés, tournant sur CPU ou sur FPGA. Point Grey propose les systèmes Bumblebee[14], à deux ou trois caméras, avec différentes options de capteurs ou de focales, et un SDK sur CPU pour le calcul des cartes de profondeur. La tête stéréo Tyzx DeepSea[15], proposée avec plusieurs options d'entraxe, utilise un FPGA et un PowerPC embarqués pour le calcul de la disparité, et transmet les données 3D par Ethernet. Focus Robotics propose la petite tête nDepth[16], avec un entraxe fixe de 6 cm, calibrée en usine et monochrome. Videre Design[17] propose des têtes stéréo à entraxe fixe ou variable, avec un calcul de la disparité par logiciel (*Small Vision System*, développé par le SRI) ou par un *chip* dédié (STOC – *Stereo On Chip*). Surveyor Corporation [18] vend le *Stereo Vision System* (SVS) qui est une solution à bas prix pour faire de la stéréo, avec possibilité de capture embarquée, motorisation, et transmission Wifi, basé sur un *firmware open source*.

---

13. La synchronisation s'effectue soit par une liaison spécifique de type *trigger* maître-esclave entre caméras, soit par le bus de transfert des images (par exemple, la plupart des caméras de la société Point Grey situées sur le même bus *firewire* se synchronisent automatiquement).

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

15. www.tyzx.com/products/cameras.html.

16. www.focusrobotics.com/.

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

18. www.surveyor.com/.

3.2.3. *Bases de données associées*

Le projet européen QUALINET[19] a recensé et classifié de nombreuses bases de données multimédia, et propose une section *3D Visual Content Databases* contenant des pointeurs vers des bases de données d'images fixes ou de vidéos stéréoscopiques ou multivues. Le projet MOBILE-3DTV[20] propose également de nombreuses séquences stéréoscopiques de référence. D'autres bases de données de qualité sont disponibles grâce au IEEE-3D *Quality Assesment Standard Group* [21], ou à l'équipe Sigmedia du Trinity College de Dublin[22].

3.3. Systèmes multivues latéraux ou directionnels

3.3.1. *Description technique*

Cette section concerne des systèmes ou dispositifs de capture multivues avec points de vue proches les uns des autres (relativement à la scène captée), souvent répartis de façon régulière sur une courbe (rectiligne ou non) ou une grille (plane ou non). On y retrouve ainsi des systèmes élaborés par assemblage mécanique (linéique ou matriciel) et synchronisation de caméras classiques comme des dispositifs construits par intégration de composants opto-électroniques placés de sorte à fournir l'arrangement désiré des points de vue puis synchronisés par une électronique dédiée. Enfin, ces outils de capture se différencient aussi par l'usage envisagé de la capture multivues (visualisation multiscopique directe, FTV, reconstruction, *refocus*, etc.) avec un impact direct sur le compromis entre nombre de vues et résolution desdites vues pour conserver une volumétrie acceptable de pixels captés, transmis et stockés.

Ces outils de capture multivues rapprochées (assemblés ou intégrés) sont souvent classifiés en *camera array* (grilles ou arrangements linéiques de caméras ou de points de vue) et systèmes ou caméras « plénoptiques ». Les *camera arrays* sont alors généralement dédiés à la capture de vues multiples de résolutions significatives pour la reconstruction de profondeur et la visualisation relief et/ou interactive (FTV) alors que les systèmes plénoptiques visent généralement la capture de « champ lumineux » (*light field*) plus équilibrés entre nombre de vues et résolution des vues pour en extraire des points de vue interpolés (FTV) ou en dériver *a posteriori* des images à mise au point variable (*refocus*) mais aussi, parfois, des reconstructions de profondeur. Cette classification est en fait plus délicate qu'il n'y paraît car la proximité de leurs géométries de capture et la croissance des capacités volumétriques de capture et traitement

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

20. http://sp.cs.tut.fi/mobile3dtv/stereo-video.

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

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

de pixels tendent à faire converger les rapports nombre de vues/nombre de pixels par vue et donc à rendre les applications visées accessibles aux deux types de systèmes. Cette classification pourrait bientôt relever d'un artefact historique lié à l'apparition par vagues successives de ces technologies et de leurs objectifs initiaux.

Indéniablement, les premiers dispositifs proposés relevaient de la classe des arrangements linéiques de points de vue. Tout d'abord limités à capturer des scènes statiques (en composition comme en éclairage), les tout premiers opéraient des prises de vue multiples par déplacement contrôlé d'une caméra comme dans la proposition de l'Université de Stanford [LEV 96]. Ils ont très vite été supplantés par des dispositifs multicapteurs saisissant plusieurs vues d'une même scène dynamique simultanément comme celui (argentique) proposé par Dayton Taylor en 1996 [TAY 96], et/ou en désynchronisation très faible et pilotée comme le système développé par Manex Entertainment pour le film « Matrix ». La plupart de ces dispositifs étaient assemblés et souvent dédiés à des applications ciblées : le projet MERL 3DTV de Mitsubishi [MAT 04], positionnait ainsi seize caméras sur un rail pour produire du contenu multiscopique destiné à ses écrans autostéréoscopiques *ad hoc* alors que l'Université de Californie à San Diego, avec Mitsubishi [JOS 06] utilisait un rail de huit caméras pour une application de détourage vidéo (*video matting*) automatique. Quelques prototypes de dispositifs intégrés ont aussi été proposés, là aussi, avec des applications ciblées. Citons notamment ceux de caméras huit points de vue développés à Reims [PRE 10], illustrés en figure 3.3, et dédiés à la production de contenus multiscopiques à déformation contrôlée (voir chapitre 4) pour des écrans autostéréoscopiques du marché.

Ces arrangements linéiques ont aussi, parallèlement, été étendus par plusieurs laboratoires en systèmes assemblés plus complexes de grilles 2D de caméras. Le plus connu est probablement celui de l'Université de Stanford[23] [WIL 05] qui a été utilisé pour de multiples applications notamment orientées vers la FTV et le *refocus*. Il est constitué d'un nombre variable de caméras (usuellement plus de cent) assemblables selon diverses configurations en grilles 2D planaires ou planaires par morceaux. Une autre grille 2D, irrégulière celle-là, a été développée à l'Université Carnegie Mellon [ZHA 04] avec 48 caméras à déplacements individuels latéraux et zénithaux contrôlés de façon à optimiser le calcul de profondeur permettant de générer la vue désirée (FTV). On peut aussi citer Sony et l'Université de Columbia qui proposent dans [NOM 07] des grilles 1D et 2D flexibles et extensibles, constituées de supports en matériau élastique sur lesquels vingt caméras sont fixées en positions régulières (au repos). La déformation du support permet alors de modifier la configuration du système pour l'adapter à la scène et au besoin exprimé (mosaïquage plus ou moins panoramique dans [NOM 07]).

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

L'émergence des grilles a aussi permis de s'intéresser à l'« espace des rayons lumineux » (*ray-space*) associé à la « fonction plénoptique » notamment synthétisée par [ADE 91]. Cette fonction plénoptique (agrégation du latin *plenus* – complet – avec optique) est la fonction qui donne l'intensité lumineuse de tous les rayons d'une scène. A valeur réelle, elle est calculée à partir de sept variables réelles : trois pour la position d'un point du rayon, deux pour sa direction 3D de propagation, une pour la longueur d'onde dont on mesure l'intensité, en enfin la dernière pour l'instant de cette mesure (en ce point) :

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

Il est usuel de réduire cette fonction à cinq variables en externalisant la longueur d'onde dans le résultat qui devient un spectre et en considérant que l'intensité est constante à l'instant de mesure sur toute la longueur du rayon[24]. Sous cette hypothèse, tous les points du rayon délivrent quasiment le même spectre au temps étudié et l'on peut donc réduire cette redondance par suppression d'une variable d'espace. En pratique, on choisit le plus souvent des points coplanaires en acceptant de ne pas « gérer » les rayons parallèles à ce plan de capture des rayons. Cela donne :

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

La dimension du domaine peut être encore réduite à quatre en figeant le temps d'étude ou en le reléguant dans le résultat qui devient alors un spectre temporel :

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

La numérisation de la fonction plénoptique réduite implique des opérations de fenêtrage puis d'échantillonnage spatial, angulaire, spectral et temporel puis de quantification des intensités qui en limite le support comme l'espace de valeurs. Ces opérations livrent une suite temporelle de signaux numériques 4D indexés par les indices i, j (liés à x, y) des points de captation situés sur une grille et les coordonnées s, t du pixel de l'image captée (en i, j), représentatives de la direction ϕ, θ du rayon mesuré en i, j, s, t. Ils contiennent pour chaque échantillon un jeu d'intensités quantifiées pour des bandes spectrales en nombre discret (généralement 3 – RGB). Ces *light fields*

24. En remarquant que l'on échantillonne temporellement à un pas dt puis que l'intensité lumineuse est transportée à la vitesse de la lumière c soit $\mathcal{I}(x,t) = \mathcal{I}(x0, t-(x-x0)/c)$, cette hypothèse est raisonnable si la dimension maximale de la scène est nettement inférieure au chemin parcouru par un photon entre deux instants d'étude, soit $299\,792\,458.dt$ m $\approx 12\,491$ km à 24 Hz, 2 998 km à 1 kHz ou encore 300 km à 1 MHz.

peuvent être obtenus aisément à partir de capture par *camera array* par simple empilement des vues captées en suivant l'ordonnancement de la grille :

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

L'attrait croissant pour cette représentation des captures multivues et, surtout, pour les modélisations et applications qui en découlent (FTV, *refocus* pour citer les principales) a permis l'arrivée d'optiques spéciales comme celle de Todor Georgiev d'Adobe-Qualcomm[25] et de solutions intégrées, les « caméras plénoptiques » proposées depuis quelques années par des sociétés comme Raytrix[26] ou Lytro[27] (voir figure 3.3). Ces caméras intégrent généralement une grille de microlentilles en amont ou en aval de l'optique de façon à capter séparément, après déviation, des rayons lumineux qui seraient sommés dans une caméra classique (voir figure 3.4 pour une illustation avec réseau lenticulaire en fond de chambre). Si l'objet capté est dans le plan de mise au point (cas B de la figure 3.4), on obtient à la place d'un pixel net, une micro-image homogène qui est synonyme de positionnement de l'objet dans le plan de focalisation. Sinon (cas A et C), on obtient plutôt qu'un pixel flou, un échantillonage local de l'objet qui, couplé à ceux des positions de capture voisines permet de reconstruire les points hors plan de focalisation. D'autres approches, notamment celle de Mitsubishi [VEE 07] [28], remplacent le réseau lenticulaire par un masque imprimé comparable aux barrières de parallaxe. Ainsi, le débat entre trous d'aiguille et microlentilles, bien connu pour les écrans stéréoscopiques, semble se transposer aux caméras plénoptiques.

Figure 3.3. *Exemples de caméras intégrées : à gauche : Cam-Box prototype de caméra huit points de vue intégrée développée par 3DTV Solutions et l'Université de Reims et, à droite : caméra plénoptique Lytro*

Pour finir, une tendance récente concerne la miniaturisation de petites grilles au sein de nouveaux composants intégrés dédiés notamment aux terminaux mobiles. La

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

26. www.raytrix.de/index.php/Cameras.html.

27. https ://www.lytro.com/camera.

28. http://web.media.mit.edu/ raskar//Mask/.

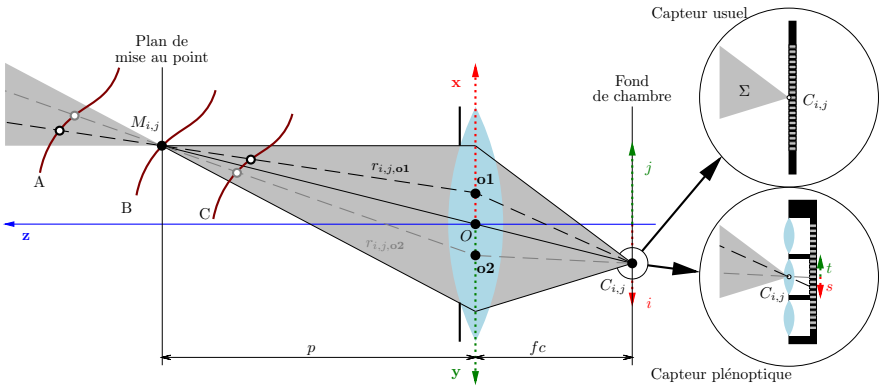


Figure 3.4. *Différences entre caméras usuelle et plénoptique : vue de dessus (axes* $\boldsymbol{x}, j, s$ *en rouge) ou de côté (axes* $\boldsymbol{y}, j, t$ *en vert) les rayons convergents en un point du fond de chambre y sont sommés dans la première, et différenciés par diffraction et échantillonnage dans la seconde*

société californienne Pelican Imaging propose ainsi un composant à microgrille 5×5 de la taille d'un capteur monovue actuel[29].

3.3.2. *Usages principaux*

Les arrangements linéiques de points de vue permettent, par simple sélection (voire interpolation) du point de vue, des effets de mouvement de caméra autour d'une scène figée ou évoluant en fort ralenti. Cette technologie, qualifiée de *bullet time* a été largement médiatisée en 1999 par le film « Matrix ». Elle est depuis proposée par plusieurs sociétés grâce à des systèmes propriétaires plus ou moins intégrés et versatiles avec des applications variées, parfois surprenantes comme le surf[30].

Avec l'émergence des dispositifs de visualisation multiscopique (voir chapitre 14), la question de la création de contenus adaptés par captation du réel a été posée et a notamment suscité plusieurs développements de *camera arrays*. Des arrangements linéiques ont ainsi été dédiés aux dispositifs autostéréoscopiques à simple parallaxe horizontale. De même, des grilles ont été proposées pour les dispositifs à double parallaxe, qualifiés d'« écrans à imagerie intégrale » (*integral imaging displays*) en référence au concept précurseur de « photographie intégrale » conjecturé [LIP 08b] puis démontré experimentalement [LIP 08a] en 1908 par Gabriel Lippmann.

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.

La génération de points de vue intermédiaires (FTV, « rendu basé image », *image based rendering*, IBR) a aussi beaucoup pesé dans l'émergence de différents *camera arrays*. Cette technologie peut apparaître comme une extension de la technique de caméra virtuelle à temps figé par interpolation de position de caméra. Sa mise en œuvre est pourtant maintenant bien différente et repose soit sur une reconstruction de profondeur permettant une reprojection des vues disponibles sur la caméra virtuelle (voir chapitre 9) soit sur une coupe par un plan (à coordonnées i, j, réelles, fixées) du champ lumineux (*light field*), signal numérique échantillonnant la fonction plenoptique réduite selon l'équation (3.10).

La forte redondance des captures multipoints de vue proches d'une même scène permet d'envisager une reconstruction de profondeur avec une robustesse accrue. La qualité des cartes de profondeur (ou de disparité dans le cas de capture en géométrie parallèle) et de la détection des occultations étant primordiale pour les applications dérivées (notamment FTV et RA), nombre d'équipes se sont penchées sur l'opportunité d'utiliser ces fortes redondances qui posent néanmoins de nouveaux défis. De multiples solutions sont ainsi proposées soit en cherchant une cohérence entre appariements binoculaires multiples, soit en cherchant directement des appareillements multi-oculaires simultanés sur toutes les vues. Quelle que soit l'approche, la gestion des occultations, accessible en vision multi-oculaire, est une aubaine qui reste délicate à gérer. Le chapitre 7 donne une description plus détaillée de ce domaine.

De façon assez comparable, l'opportunité de disposer de vues fortement redondantes permettant un appareillement global est utilisée (voir chapitre 19) pour proposer des dispositifs de capture à grande gamme dynamique (HDR) recalculée à partir des captures à gammes dynamiques modérées mais variées des différents points de vue. La variation des gammes selon les points de vue est obtenue par des filtres neutres de densité différentes ou par des réglages distincts du temps d'exposition.

Pour finir, nous évoquerons un usage des captures multivues par grille ou caméras plénoptiques qui peut surprendre tant la notion de profondeur de champ, cruciale en photographie, semblait immuablement réglée à la prise de vue. Les captures à grand nombre de vues comme la modélisation en *ray-space* ont fait éclore une activité foisonnante autour d'une opportunité nouvelle aux implications très prometteuses : le choix *a posteriori* de la mise au point (*refocus*). Cela inclut notamment :

– la sélection du plan de mise au point (par moyenne des pixels de plusieurs vues correspondant aux rayons provenant géométriquement de mêmes points de ce plan) ;

– le choix de l'ouverture donc de la profondeur de champ (par sélection du voisinage de points de vue dont sont tirés les pixels moyennés) ;

– la possibilité de choisir une profondeur de champ infinie – *all in focus* – (par sélection de pixels non moyennés, ce qui correspond à l'ouverture fine d'un sténopé) ;

– la suppression de premier plan de certaines images pour y exhiber l'arrière-plan caché s'il est assez lointain pour être visible depuis d'autres points de vue.

3.3.3. *Bases de données associées*

Sans chercher à être exhaustif, on peut citer quelques sites de données acquises par des dispositifs relevant de cette section. L'Université de Californie à San Diego et Mitsubishi proposent[31] des captures en arrangement linèique : vidéos 8-vues et séries de 120 à 500 images statiques. La bibliothèque de *light fields* de l'Université de Stanford[32] est riche de multiples scènes très variées capturées en haute résolution souvent depuis plusieurs centaines de points de vue, notamment par déplacement de caméra sur bras robotisé ou par la grille de Stanford. Ces données sont disponibles en version brute ou rectifiée, avec données de calibration et possibilité d'interaction en ligne avec leur forme *light field* par sélection de point de vue et manipulation de *refocus* (choix d'ouverture et de plan de mise au point). Cette bibliothèque complète et supplante la version précédente[33] qui propose des séries plus légères tant en nombre de vues qu'en résolution. Dans une mesure plus modeste, Todor Georgiev livre sur son site[34], quelques images plénoptiques de plusieurs dizaines de millions de rayons parfois. Enfin, l'Université d'Heidelberg maintient aussi une bibliothèque[35] de quelques *light fields* de synthèse, alors fournis avec information de profondeur « terrain », mais aussi capturés du réel par caméras plénoptiques Raytrix, tous avec une grille 9×9.

3.4. Systèmes multivues englobants ou omnidirectionnels

3.4.1. *Description technique*

Nous abordons ici les systèmes de capture multicaméras réparties de façon assez espacée et approximativement convergente de sorte à « couvrir » avec suffisamment de redondance un espace scénique assez grand pour y faire évoluer les objets et/ou acteurs. Les premiers systèmes de ce genre ont été déployés pour des techniques de *bullet time* ou de *motion capture*. Les systèmes « englobant » utilisés pour le temps figé sont généralement constitués d'un rail formant une courbe représentant la trajectoire désirée pour la caméra virtuelle (fermée ou non, pas toujours plane ou circulaire, etc.) portant des caméras en nombre souvent important à visée réglée selon celle désirée pour la caméra virtuelle à cet endroit et avec une synchronisation pilotée dépendant de l'effet désiré (temps figé ou plus ou moins ralenti). En MoCap par vidéo avec marqueurs, on utilise majoritairement un nombre plus réduit de caméras infrarouges synchronisées en disposition libre avec un processus de calibration géométrique par déplacement d'un objet mire à marqueurs fixes.

---

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.

L'usage assez intensif de ces techniques par les industries du cinéma et du jeu vidéo (en mesure de le rentabiliser), a suscité un intérêt marqué pour une technologie plus aboutie utilisant des captures multivues sans marqueur et délivrant des résultats d'usages plus variés : la « vidéo 3D ». Proposée dès 1997 [KAN 97, MOE 97] et intensément étudiée et développée depuis lors [MAT 12], elle permet de reconstruire tout au long de la séquence d'acquisition la géométrie comme l'apparence (texture) de l'objet ou acteur filmé pour en enregistrer un avatar numérique animé de qualité suffisante pour être réutilisé par synthèse d'image sous des angles de vue très peu restreints.

Cela nécessite un système de capture multivues synchronisées avec de nombreux points de vue répartis autour de l'espace scénique utile caractérisé comme intersection des zones de profondeur de champ des caméras (voir gauche de la figure 3.5). Le compromis entre nombre de caméras (complétude) et écart entre caméras (précision de reconstruction) a été posé par [KAN 97] entre neuf et seize caméras pour une disposition régulière sur un cercle à mi-hauteur de la scène avec convergence au centre du cercle (voir haut gauche de la figure 3.5 pour un exemple à douze caméras). Des solutions plus complètes ont ensuite été proposées pour accéder à la reconstruction des sommets des objets en ajoutant des caméras en orientation plongeante au-dessus de la scène, puis en choisissant des dispositions échantillonnant plus régulièrement les directions de capture (plusieurs cercles à hauteurs différentes avec caméra(s) zénithale(s)[36], dômes[37] [38] ou en arrangements plus libres en studio ou en extérieur [KIM 12][39]) avec un nombre de caméras fluctuant selon les contextes applicatifs de quelques unités (*University of Surrey*[39], *Max Planck Institute* [AGU 08] ou projet « GrImage »[40]) à plusieurs centaines (1 000 pour le projet « Virtualized reality »[41]).

Ces assemblages complexes doivent aussi être dotés de capacités conséquentes en termes de réseau, stockage et calcul pour gérer les flux vidéo générés et de technologies de calibration géométrique comme colorimétrique très précises. Enfin, maîtriser l'éclairage et simplifier le détourage des objets facilite le traitement des images. Tout cela rend ces systèmes complexes, délicats et onéreux et explique leur organisation usuelle en salles dédiées parfois appelées « studios vidéo 3D ».

---

36. Recover3D, projet « investissements d'avenir », 2012-2014, piloté par XD Productions, voir haut droit et bas de la figure 3.5.

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

38. Projet 3D-COFORM FP7 2007-2013, www.vcc-3d.eu/multiview et www.3dcoform.eu, numérisation de patrimoine pour de petits objets à interactions lumière/matière complexes.

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

40. www.inrialpes.fr/grimage/.

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

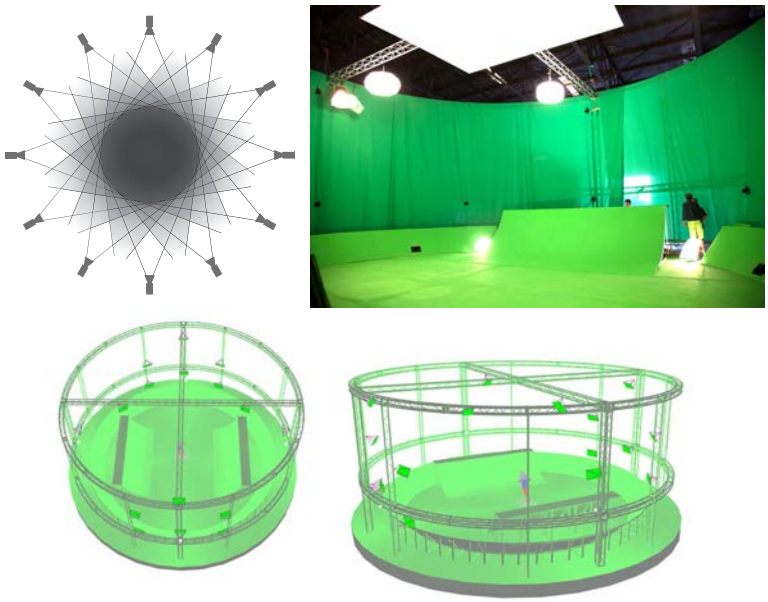


Figure 3.5. *Exemples de studios vidéo 3D : en haut à gauche, schéma d'arrangement circulaire de douze caméras montrant l'espace scénique utile comme intersection des zones de profondeur de champ des caméras (en gris transparent) ; en haut à droite et en bas studio du projet Recover3D*[36]

Le marché du *bullet time* est principalement structuré autour de sociétés de services[42], qui opérent des systèmes propriétaires alors que celui de la MoCap est aussi occupé par plusieurs acteurs[43], qui distribuent des solutions « clés en main ». Pour ce qui est de la vidéo 3D, l'offre de service se développe avec des sociétés de production spécialisées dotées de studios 3D[44] alors que celle de la commercialisation de systèmes semble démarrer[45].

42. Citons par exemple Reel EFX www.reelefx.com/ et Time Slice www.timeslicefilms.com/#1.

43. Citons notamment Vicon (www.vicon.com/), Animazoo (www.animazoo.com/) et Moven (www.moven.com/).

44. Citons XD Productions (www.xdprod.com/) et 4D View Solutions (www.4dviews.com/).

45. 4D View Solutions www.4dviews.com/ commercialise aussi des solutions depuis quelques temps.

3.4.2. *Usages principaux*

Nous parlerons peu ici des technologies de temps figé ou de MoCap car leurs systèmes de captation assez particuliers les positionnent à la limite du champ de cet ouvrage. Ainsi, le principal usage des systèmes que nous qualifions d'englobant réside dans la vidéo 3D, en plein essor tant en recherche qu'en production, comme en atteste [MAT 12], ouvrage totalement dédié à cette technique. La vidéo 3D s'appuie sur des systèmes complexes, incluant de nombreuses caméras réparties, synchronisées et calibrées en géométrie et colorimétrie, puis un réseau de transfert des flux vidéo, et enfin des capacités de calcul et de stockage assez conséquentes.

L'extraction de la géométrie des avatars depuis les flux vidéo multiples nécessite pour commencer une calibration géométrique très précise de toutes les caméras. Cette reconstruction peut être opérée selon trois techniques classées en méthodes « basées modèle » ou, par opposition, méthodes libres. La première classe correspond à la recherche de la configuration d'un modèle prédéfini qui optimise les degrés de liberté du modèle géométrique cherché pour que ses projections correspondent au mieux aux images captées. La seconde contient deux techniques concurrentes : la multistéréovision qui cherche à reconstruire des points 3D par triangulation à partir des pixels jugés homologues dans des images différentes et les méthodes « basées silhouettes » qui reconstruisent l'enveloppe visuelle de l'avatar par intersection des cônes généralisés que forment ses projections détourées dans toutes les images. La recherche de modèle prédéfini souffre par construction d'un défaut très souvent rédhibitoire : le manque d'adaptabilité, elle peut néanmoins guider une reconstruction par silhouettes avec moins de caméras ([AGU 08], projet « Free Viewpoint Video of Human Actors »[46] [CAR 03]). Les méthodes de stéréovision sont sensibles aux erreurs de calibration colorimétrique comme aux phénomènes spéculaires, généralement assez lourdes en temps de calcul mais sont en mesure de délivrer des détails géométriques dans les zones concaves là où l'enveloppe visuelle resterait par nature convexe. Inversement, les enveloppes visuelles sont plus aisées à obtenir, à moindre coût, par des méthodes, dites « *Visual Hull* », plus robustes mais ces enveloppes délivreront par essence des résultats grossiers dans les zones concaves des objets. Les techniques basées modèles sont assez souvent employées pour la numérisation d'acteurs humains. Dans le contexte libre, même appliqué à des humains, celles de *Visual Hull* (objet du chapitre 8) sont plus souvent retenues en production pour leur robustesse mais leurs limitations freinent aujourd'hui leur progression. C'est pourquoi la complémentarité entre multistéréovision et silhouettes a suscité des projets basés sur leur hybridation comme Recover3D[36] qui propose de répartir autour de l'espace scénique des caméras monoscopiques et des caméras multiscopiques pour produire un modèle géométrique robuste (par intégration dans l'enveloppe visuelle) et plus détaillé (par reconstruction multistéréoscopique) notamment dans les parties concaves.

---

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

Une fois le modèle géométrique reconstruit à chaque pas de temps, il reste à lui attribuer à partir des images captées un contenu visuel (une texture) cohérent dans le temps. On applique alors des solutions de suivi temporel des modèles géométriques (voir chapitre 8) pour assurer la cohérence sémantique des accroches des textures, puis des techniques de *video-texturing* qui consistent à mixer localement les informations photométriques reprojetées sur le modèle géométrique depuis les images où cette zone locale n'est pas occultée. Les difficultés tiennent ici aux choix à opérer quand on constate des écarts entre les données rétroprojetées. Ces écarts peuvent venir de défauts de reconstruction géométrique, de défaut de calibration colorimétrique comme de caractéristiques liées à la scène elle-même comme des reflets ou autres phénomènes spéculaires. Ces phénomènes optiques complexes sont à la base de projets dédiés comme la série des *light stages*[47], systèmes dédiés à la capture de propriétés optiques complexes dans un contexte de *camera array* avec modulation de l'éclairage ou, plus récemment, le projet 3D-COFORM[38] qui ambitionne la numérisation de haute qualité d'objets patrimoniaux et culturels par acquisitions d'objets statiques sous de multiples conditions d'éclairage (151 sources) depuis 151 points de vue et sous différentes expositions pour en déduire des vues HDR (une par couple source/point de vue) permettant alors un plaquage de propriétés optiques sous forme de fonctions de textures bidirectionnelles (BFT).

La vidéo3D est plus onéreuse à la capture que la MoCap car plus complexe. Cependant, l'usage de ses résultats est beaucoup plus versatile. En effet, le réalisateur avec ses graphistes peut, en postproduction, aisément choisir ses angles de vue avec peu de limites spatiales tout en éditant les avatars animés acquis dans ses scènes (déplacement/déformation spatio-temporels, duplication, transposition dans d'autres scènes, rééclairage[48], etc.). Ces possibilités permettent de mieux rentabiliser les avatars acquis et donc de diminuer les coûts de production. Cela donne une technologie à la fois plus ouverte à la créativité et plus économique qui, par conséquent, devient accessible à la production télévisuelle. Par ailleurs, cette numérisation d'avatars animés intéresse aussi d'autres domaines applicatifs comme la culture[38], le sport [KIM 12] ou la téléprésence collaborative [PET 10].

Enfin une tendance récente, hors du sujet de ce chapitre, extrapole les attendus de la vidéo 3D ci-dessus décrits : la reconstruction 3D à partir de sources collectives non calibrées (par exemple des captures d'amateurs trouvées sur le *web*) sous forme de photos [GOE 07, SNA 09] ou de vidéos ([BAL 10], projet « Virtual Video Camera »[49]).

---

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

48. Le lecteur pourra en trouver quelques illustrations sur le site de XD Productions www.xdprod.com/Xd Productions_RD.swf.

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

3.4.3. *Bases de données associées*

Quelques sites académiques mettent à disposition de la communauté des séquences multivues captées par leur système : l'Université du Surrey livre ainsi quelques captures 8-vues en arrangement circulaire (www.ee.surrey.ac.uk/cvssp/visualmedia/visualcontentproduction/projects/surfcap), le MIT propose une dizaine de jeux de données complets (images, poses, résultats, etc.) captés et traités selon [VLA 08] (http://people.csail.mit.edu/drdaniel/mesh_animation/) et l'INRIA Rhône-Alpes offre sur son « 4D repository » quelques dizaines de jeux captés par leurs systèmes GrImage et IXmas (http://4drepository.inrialpes.fr/).

3.5. Conclusion

Ce chapitre a montré que la capture multivues recouvre des technologies variées, toutes complexes. Ces technologies qui ouvrent la voie à des postproductions plus créatives pourraient révolutionner la logique de production audiovisuelle en offrant plus de possibilités de retravail qualitatif des médias capturés *a posteriori* de la prise de vue. Elles ouvrent aussi la voie à une numérisation toujours plus riche de notre environnement comme à nombre d'autres domaines applicatifs nécessitant reconstruction 3D et/ou capture ou reconnaissance de mouvements. Si ces technologies se matérialisent principalement à ce jour dans des prototypes de laboratoire, des systèmes *ad hoc* de sociétés de service ou des dispositifs de petite série, l'importance de toutes ces applications devrait en permettre le développement commercial comme semble l'attester l'arrivée des caméras plénoptiques et des microgrilles pour terminaux mobiles.

3.6. Bibliographie

[ADE 91] ADELSON E. H., BERGEN J. R., « The Plenoptic Function and the Elements of Early Vision », *Computational Models of Visual Processing*, p. 3–20, MIT Press, Cambridge, MA, Etats-Unis, 1991.

[AGU 08] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S., « Performance capture from sparse multi-view video », *Proceedings ACM SIGGRAPH*, vol. 27, Los Angeles, CA, Etats-Unis, p. 98 :1–98 :10, août 2008.

[BAL 10] BALLAN L., BROSTOW G. J., PUWEIN J., POLLEFEYS M., « Unstructured video-based rendering : interactive exploration of casually captured videos », *Proceedings ACM SIGGRAPH*, Los Angeles, CA, Etats-Unis, p. 87 :1–87 :11, juillet 2010.

[CAR 03] CARRANZA J., THEOBALT C., MAGNOR M. A., SEIDEL H.-P., « Free-viewpoint video of human actors », *Proceedings ACM SIGGRAPH*, San Diego, CA, Etats-Unis, p. 569–577, juillet 2003.

[DEV 10] DEVERNAY F., BEARDSLEY P., « Stereoscopic Cinema », RONFARD R., TAUBIN G., Eds., *Image and Geometry Processing for 3-D Cinematography*, vol. 5 de *Geometry and Computing*, Chapitre 2, p. 11–51, Springer, Heidelberg, Allemagne, 2010.

[EMO 05] EMOTO M., NIIDA T., OKANO F., « Repeated Vergence Adaptation Causes the Decline of Visual Functions in Watching Stereoscopic Television », *Journal of Display Technology*, vol. 1, n°2, p. 328–340, décembre 2005.

[GOE 07] GOESELE M., SNAVELY N., CURLESS B., HOPPE H., SEITZ S. M., « Multi-View Stereo for Community Photo Collections », *Proceedings ICCV, IEEE International Conference on Computer Vision*, Rio de Janeiro, Bresil, p. 1–8, octobre 2007.

[JOS 06] JOSHI N., MATUSIK W., AVIDAN S., « Natural Video Matting using Camera Arrays », *Proceedings ACM SIGGRAPH*, vol. 25, Boston, MA, Etats-Unis, p. 779–786, juillet 2006.

[KAN 97] KANADE T., RANDER P., NARAYANAN P. J., « Virtualized Reality : Constructing Virtual Worlds from Real Scenes », *IEEE MultiMedia*, vol. 4, n°1, p. 34–47, IEEE Computer Society Press, janvier 1997.

[KIM 12] KIM H., GUILLEMAUT J.-Y., TAKAI T., SARIM M., HILTON A., « Outdoor Dynamic 3-D Scene Reconstruction », *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, n°11, p. 1611–1622, novembre 2012.

[LEV 96] LEVOY M., HANRAHAN P., « Light field rendering », *Proceedings ACM SIGGRAPH*, Nouvelle Orléans, LA, Etats-Unis, p. 31–42, août 1996.

[LIP 08a] LIPPMANN M. G., « Epreuves réversibles donnant la sensation du relief », *Journal de Physique Théorique et Appliquée*, vol. 7, n°1, p. 821–825, novembre 1908.

[LIP 08b] LIPPMANN M. G., « Epreuves réversibles. Photographies intégrales », *Comptes Rendus de l'Académie des Sciences*, vol. 146, n°9, p. 446–451, mars 1908.

[LIP 82] LIPTON L., *Foundations of the Stereoscopic Cinema*, Van Nostrand Reinhold, New York, NY, Etats-Unis, 1982.

[MAR 99] MARCOS S., MORENO E., NAVARRO R., « The depth-of-field of the human eye from objective and subjective measurements », *Vision Research*, vol. 39, n°12, p. 2039–2049, juin 1999.

[MAT 04] MATUSIK W., PFISTER H., « 3D TV : A Scalable System for Real-Time Acquisition, Transmission, and Autostereoscopic Display of Dynamic Scenes », *Proceedings ACM SIGGRAPH*, vol. 24, Los Angeles, CA, Etats-Unis, p. 814–824, août 2004.

[MAT 12] MATSUYAMA T., NOBUHARA S., TAKAI T., *3D Video and Its Applications*, SpringerLink : Bücher, Springer, Londres, Royaume-Uni, 2012.

[MEN 09] MENDIBURU B., *3D Movie Making : Stereoscopic Digital Cinema from Script to Screen*, Focal Press, Burlington, MA, Etats-Unis, 2009.

[MEN 11] MENDIBURU B., *3D TV and 3D Cinema : Tools and Processes for Creative Stereoscopy*, Focal Press, Waltham, MA, Etats-Unis, 1e édition, 2011.

[MOE 97] MOEZZI S., TAI L.-C., GERARD P., « Virtual View Generation for 3D Digital Video », *IEEE MultiMedia*, vol. 4, n°1, p. 18–26, IEEE Computer Society Press, janvier 1997.

[NOM 07] NOMURA Y., ZHANG L., NAYAR S., « Scene Collages and Flexible Camera Arrays », *Proceedings EGSR, Eurographics Symposium on Rendering*, juin 2007.

[PET 10] PETIT B., DUPEUX T., BOSSAVIT B., LEGAUX J., RAFFIN B., MELIN E., FRANCO J.-S., ASSENMACHER I., BOYER E., « A 3d data intensive tele-immersive grid », *Proceedings MM, international conference on Multimedia*, Florence, Italie, ACM, New York, NY, Etats-Unis, p. 1315–1318, 2010.

[PRE 10] PREVOTEAU J., CHALENÇON-PIOTIN S., DEBONS D., LUCAS L., REMION Y., « 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*, vol. 2010, p. 1–11, Hindawi, mars 2010.

[SNA 09] SNAVELY K. N., Scene reconstruction and visualization from internet photo collections, PhD thesis, Seattle, WA, Etats-Unis, 2009.

[TAY 96] TAYLOR D., « Virtual camera movement : The way of the future ? », *American Cinematographer*, vol. 77, n°9, p. 93–100, 1996.

[UKA 07] UKAI K., HOWARTH P. A., « Visual fatigue caused by viewing stereoscopic motion images : Background, theories, and observations », *Displays*, vol. 29, n°2, p. 106–116, mars 2007.

[VEE 07] VEERARAGHAVAN A., RASKAR R., AGRAWAL A., MOHAN A., TUMBLIN J., « Dappled Photography : Mask Enhanced Cameras for Heterodyned Light Fields and Coded Aperture Refocusing », *Proceedings ACM SIGGRAPH*, vol. 26, San Diego, CA, Etats-Unis, juillet 2007.

[VLA 08] VLASIC D., BARAN I., MATUSIK W., POPOVIĆ J., « Articulated mesh animation from multi-view silhouettes », *Proceedings ACM SIGGRAPH*, vol. 27, Los Angeles, CA, Etats-Unis, p. 97 :1–97 :9, août 2008.

[WIL 05] WILBURN B., JOSHI N., VAISH V., TALVALA E.-V., ANTUNEZ E., BARTH A., ADAMS A., HOROWITZ M., LEVOY M., « High performance imaging using large camera arrays », *Proceedings ACM SIGGRAPH*, Los Angeles, CA, Etats-Unis, p. 765–776, juillet 2005.

[YAN 04] YANO S., EMOTO M., MITSUHASHI T., « Two factors in visual fatigue caused by stereoscopic HDTV images », *Displays*, p. 141–150, novembre 2004.

[ZHA 04] ZHANG C., CHEN T., « A Self-Reconfigurable Camera Array », *Proceedings EGSR, Eurographics Workshop on Rendering Techniques*, Norköping, Suéde, Eurographics Association, Aire-la-Ville, Suisse, p. 243–254, juin 2004.