Universitat Politècnica de Catalunya

Tesi Doctoral

Contribució als mètodes d'obtenció i representació de vistes d'objectes reals per aplicacions interactives

Enric Xavier Martín i Rull

Dirigida per Antonio Benito Martínez Velasco

Programa de doctorat en Control, Visió i Robòtica

Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial

Barcelona, Setembre de 2006.

Als meus pares, la Marta, en Francesc i en Miquel,

Agraïments

En primer lloc, he d'agrair al meu mestre i amic Antonio la direcció realitzada d'aquesta tesi i en general, del meu treball i aprenentatge al departament des del meu temps de becari. També vull manifestar el meu sincer agraïment a la resta de companys del departament d'ESAII de la UPC. Sense els seus consells, ànims, ajudes, mestratge, suport en la intendència i la burocràcia, converses estimulants i amistat no hauria estat possible tirar endavant aquest treball. A tots, moltes gràcies.

Glossari

Donat que aquest treball es centra en l'àmbit tecnològic, on els neologismes, anglicismes i acrònims envaeixen el llenguatge, s'ha considerat necessari crear un petit glossari amb alguns dels termes emprats. Aquest glossari mostra, pels mots seleccionats, una petita explicació, la gènesi dels acrònims i pels anglicismes una traducció a la llengua catalana[1].

Aparellament estèreo, és un procediment pel qual, donades dues imatges d'una mateixa escena preses des de diferents punts de vista, s'intenta trobar per a cada punt d'una imatge la seva projecció en l'altra. Com a resultat es pot generar un mapa de correspondències (punt a punt), un mapa de disparitat (distància d'un punt al seu corresponent en l'altra imatge, sols aplicable si les imatges són coplanars i estan alineades) o si es vol calcular la profunditat dels punts aparellats a l'espai (per això caldrà triangular la seva posició coneixent els paràmetres intrínsecs i extrínsecs de les càmeres), el resultat pot ser una superfície.

Background, terme anglès referit a la part d'una imatge que no té interès per al procés realitzat. Es pot traduir simplement com a fons de la imatge.

Baseline, terme anglès referit a la línia que, en una configuració esteroscòpica on dues càmeres observen una escena, uneix els centres òptics de les dues càmeres.

Càmera virtual, en aplicacions de síntesi de vistes, dispositiu imaginari on es projecta una escena. La imatge que suposadament captura la càmera virtual s'obtindrà per interpolació de vistes obtingudes de càmeres reals, projecció de models tridimensionals o altres tècniques similars.

CCD, acrònim de *Charge-coupled device*, literalment dispositiu de càrregues acoblades. És un sensor d'imatges consistent en una matriu d'elements fotosensors i condensadors que emmagatzemen les càrregues. Són els captadors emprats en les càmeres de fotografia i televisió. Malgrat mantenir-se el nom, els darrers anys la implementació dels captadors tendeix a fer-se amb la tecnologia de fabricació de transistors CMOS, que tot i oferir pitjors prestacions en sensibilitat a la llum, és més barata i consumeix menys energia.

[1] Per les traduccions consolidades s'han consultat les fonts [termcat 06][iec 06][grec 06]

Chroma key, tècnica emprada en visió per ordinador i en els estudis de televisió per separar fàcilment un o més objectes del fons de la imatge. Consisteix en pintar el fons amb un color no present en cap dels objectes o persones i triar si un punt de la imatge és rellevant o no en funció d'aquest color. Es pot traduir com a selecció per color.

CISC, acrònim de *Complex Instruction Set Computer*, traduïble com a ordinador amb un conjunt complex d'instruccions, en contraposició als ordinadors RISC *Reduced Instruction Set Computer*, que tenen un conjunt reduït d'instruccions. Es reflexa així una disjuntiva entre dues filosofies en el disseny dels processadors: dotar-los de poques instruccions senzilles que poden executar-se ràpidament o de moltes instruccions capaces d'executar operacions de més alt nivell però més lentament.

Convex Hull, traduïble com a carcassa convexa. És un volum imaginari mínim que conté un objecte, però sense perfilar-ne les concavitats o forats.

CPU, acrònim de *Central Processing Unit*, traduïble com a unitat central de procés. És l'element que en un ordinador o microcontrolador s'encarrega dels càlculs aritmètics i lògics i del control del flux de programa.

DSP, acrònim de *Digital Signal Processor*, o processadors digitals de senyals. Són una família de les CPU especialitzada en càlculs vectorials on una mateixa instrucció pot ser repetidament executada damunt grans conjunts de dades.

Estereovisió, disciplina que tracta la problemàtica d'extreure informació tridimensional d'una escena a partir de les vistes obtingudes per un parell o més de càmeres.

FFT, acrònim de *Fast Fourier Transform* o transformada ràpida de Fourier. És un procediment matemàtic mitjançant el qual un senyal es converteix del domini temporal al domini freqüencial. La FFT és una implementació eficient de la transformada de Fourier on els senyals d'entrada no prenen valors continus sinó discrets.

Foreground, terme anglès referit a la part d'una imatge que si que té interès per al procés realitzat, en contraposició al *background* o fons. En fotografia es tradueix com a primer pla.

FPGA, acrònim de *Field-Programmable Gate Array*, traduïble com a matriu programable de portes lògiques. És un dispositiu electrònic reconfigurable que pot realitzar massivament operacions lògiques o de control i seqüenciació de processos.

GPU, acrònim de *Graphic Processing Unit*, és a dir, unitat de procés gràfic. En els ordinadors actuals, el perifèric encarregat de la representació d'informació en pantalla (veure VGA), no va només equipat amb la memòria gràfica sinó que també porta un processador per fer la projecció d'estructures tridimensionals, il·luminació d'aquestes, o descompressió de vídeo.

HMD, acrònim de *Head Mounted Display* o pantalla muntada en el cap de l'usuari. És un dispositiu emprat sobretot en realitat augmentada o realitat virtual per representar gràfics (sovint estereoscòpics) davant dels ulls de l'usuari, aconseguint una major immersió d'aquest en l'entorn.

MMX, acrònim de les *MultiMedia eXtensions*, o extensions multimèdia, que són un conjunt d'instruccions i maquinari especialitzat per la seva execució, que en els processadors d'arquitectura CISC ofereixen suport pel tractament de senyals d'àudio i vídeo.

OpenGL paraula derivada de *Open Graphic Library*, o llibreria de funcions gràfiques oberta (en el sentit de pública), és un conjunt d'instruccions d'alt nivell per a programar gràfics tridimensionals en les GPU. Existeixen versions registrades com les *DirectX* de Microsoft.

Píxels, traducció acceptada de l'anglès *pixel* és la unitat mínima de discretització d'una imatge o dispositiu de representació d'imatges. Normalment la resolució de les imatges o d'un monitor s'expressa en aquesta unitat.

Paràmetres intrínsecs, en el càlcul del model de projecció d'una escena en una càmera, són aquells paràmetres referits als elements constructius de la càmera: distància focal, mida de l'element captador o CCD, inclinació de l'element captador respecte a l'eix òptic, distorsió de la lent i altres.

Paràmetres extrínsecs, són aquells que no depenen de l'estructura interna de la càmera sinó de la posició a l'espai i la seva orientació (referida normalment en funció de tres angles. En català s'anomenen capcineig, balanceig i guinyada)

Projector gnomònic, nom donat al sistema robotitzat d'adquisició construït per la fase experimental d'aquesta tesi. L'adjectiu gnomònic s'aplica en cartografia al sistema de projecció en que l'origen és el centre d'una esfera i el pla de projecció el pla tangent a l'esfera en un punt. Homòlogament, el sistema que s'ha dissenyat consta d'una càmera que es va movent per la superfície d'una esfera mantenint l'objectiu en el centre d'aquesta.

Realitat augmentada, disciplina que tracta l'afegit d'informació virtual (en sentit ampli imatges, sons, dades) damunt de la captada del món real per una persona. Perquè aquesta sigui útil la informació haurà d'estar en correspondència amb la del món real i s'haurà de disposar d'algun dispositiu de representació.

Registration, traduïble com a posada en correspondència. Procediment pel qual la informació del món real donada per un sensor és correlada amb la informació tinguda en un model, per efectuar-hi una superposició o barreja que serà presentada a un humà en una aplicació de realitat augmentada.

See-through glasses, literalment: ulleres que deixen veure a través, és un dispositiu emprat en aplicacions de realitat augmentada. A diferència dels HMD, les pantalles emprades per dibuixar la informació són semitransparents i deixen

veure a l'usuari l'entorn on es troba. Caldrà doncs que la informació afegida es trobi perfectament alineada amb l'escena vista pel portador del dispositiu.

Segmentació: algoritme que donada una imatge, separa la part d'interès per al procés a realitzar o *foreground* del fons de l'escena o *background*.

VGA, acrònim de *Video Graphic Adapter*, traduïble com a adaptador de sortida de vídeo. És el perifèric dels ordinadors personals que permet connectar un monitor o televisor per la visualització d'informació per als humans.

Vòxel, traducció acceptada de l'anglès *voxel*. En informàtica és la unitat mínima de discretització de l'espai. Normalment es defineix aquest espai com a Euclidià i s'hi construeix una matriu tridimensional de vòxels de resolució donada.

Índex

Índex de figures

Capítol 1

Capítol 2

Capítol 3

Capítol 4

Capítol 5

Capítol 6

Capítol 7

Capítol 8

Capítol 9

Índex de taules

Capítol 2

Capítol 3

Capítol 6

Capítol 8

Capítol 9

1. Introducció

El món actual està dominat per la imatge, en l'anomenada societat de consum la imatge d'un producte és tan o més important que la qualitat del producte en si.

Així doncs, quan es vol transmetre una idea, concepte o informació, el mitjà audiovisual s'ha imposat als altres mitjans gràcies a la immediatesa i presumpta credibilitat que té la imatge. En el camp de l'educació, no sense controvèrsia, la utilització de la imatge va substituint la descripció escrita. En les comunicacions interpersonals, s'ha vist aparèixer els darrers anys la telefonia mòbil amb transmissió de vídeo i els programes de conversa per la xarxa que poden intercanviar text, so i vídeo a temps real. En aquests camps i en el de l'enginyeria, les noves eines tendeixen a oferir un entorn de percepció cada cop més immersiu, de manera que la persona pugui comprendre i interactuar de forma natural amb un entorn immediat o distant.

Aquesta immersió de la persona en la informació, moltes vegades va més enllà del que els sentits podrien percebre per ells mateixos i s'intenta augmentar la capacitat de percepció de l'individu. Uns binocles són un exemple senzill d'augment de la capacitat perceptiva d'una persona, un aparell per a la sordesa o unes ulleres graduades també, però aquests tres exemples l'únic que fan és corregir o potenciar uns sentits dels que l'individu disposa per naturalesa. La tecnologia ha permès també afegir informació elaborada a allò que estem percebent. Aquesta informació, no la tindria la persona de forma natural, sinó que hauria d'utilitzar diferents aparells de mesura per obtenir-la, i crear una associació mental del que està percebent amb les dades suplementàries. Quan es presenta a una persona informació convencional del món real, on s'hi afegeixen dades més o menys elaborades, parlem de que s'està fent realitat augmentada. Un exemple força conegut de realitat augmentada és la projecció als vidres de la cabina d'un avió, de la informació captada pel radar. Abans de desenvolupar aquest sistema, el pilot havia de mirar alternativament el paisatge natural i la pantalla del radar en el quadre de comandament. Volant a altes velocitats i realitzant maniobres arriscades, això li feia perdre uns segons que podien ser crítics.

La realitat augmentada no té massa sentit si la informació afegida no està lligada d'alguna manera amb la percepció natural de l'individu. Si el que s'afegeix no està correlat en l'espai i en el temps amb el que la persona veu, serà inútil. De res serviria al pilot de l'avió que se li mostrés la informació del radar de fa cinc segons o se li mostrés girada 45 graus. Cal doncs establir una relació precisa entre les dades naturals, en l'exemple imatges, i les dades artificials, en l'exemple informació del radar. L'establiment d'aquest lligam entre els dos móns, s'anomena en anglès *registration* i es podria traduir aproximadament com a correspondència. Es presentarà ara un altre exemple de realitat augmentada: moltes vegades a la nit, mirant els estels, s'aprecien les

formes de les constel·lacions. Un observador no expert desconeixerà el nom d'algunes d'elles, la seva disposició o el nom de les estrelles i altres cossos celestes. Un sistema de realitat augmentada, podria donar una imatge com la següent:

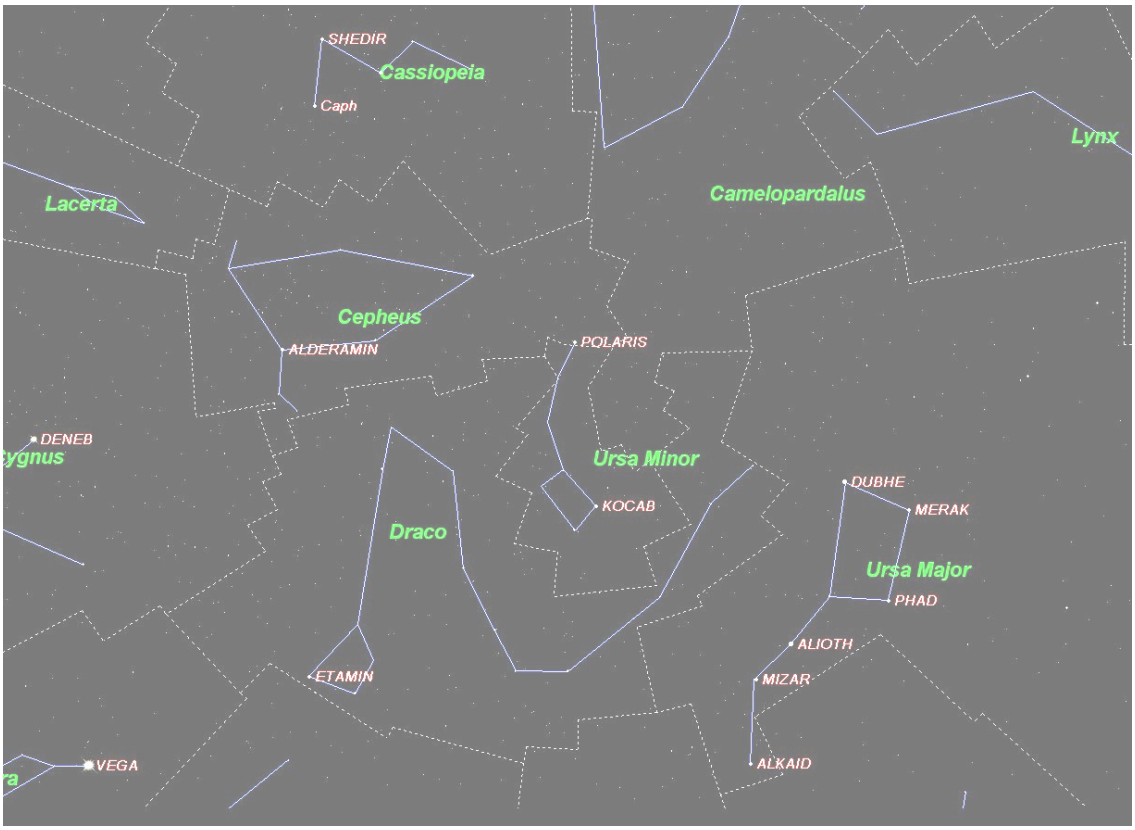
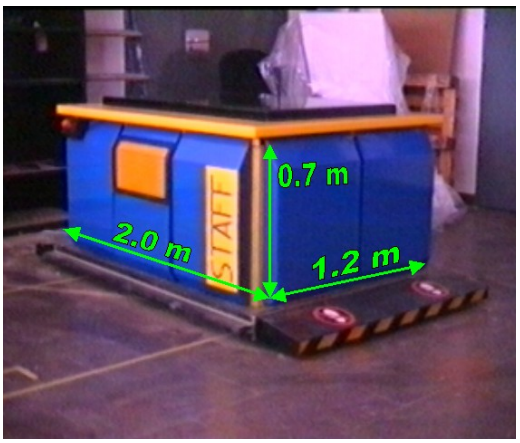


Figura 1.1. Imatge celeste etiquetada amb realitat augmentada

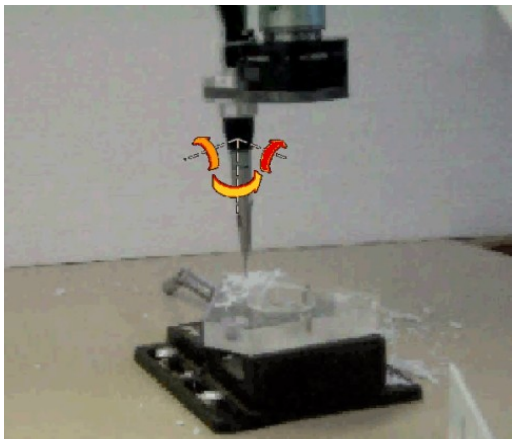
Què li caldria a aquest sistema per funcionar? dues coses bàsiques: una senzilla, que és una base de dades dels cossos celestes i una complicada, que és determinar la posició i orientació exacta dels ulls de l'observador del cel. Si es tinguessin aquests dos elements es podria fer la correlació de la imatge captada pels ulls amb la informació de la base de dades, projectada també en forma d'imatge, és a dir, la correspondència. Finalment, caldria un dispositiu com unes *see-through glasses*[1] que òpticament barreges les dues imatges i les presentés.

En funció de l'aplicació, existeixen diferents maneres de construir i mostrar la realitat augmentada. A vegades interessa presentar informació afegida numèrica, textual, en forma de línies, de canvis de coloració o d'objectes no presents a l'escena. Quan el que es tracta és una seqüència d'imatges apareixen les restriccions del temps de procés per la correspondència i barreja de la informació. Caldrà realitzar aquestes dues tasques en un temps determinat, de l'ordre de la vintena de milisegons. A continuació es mostren uns exemples de realitat augmentada, realitzats tots durant el desenvolupament d'aquesta tesi: usant línies i nombres per a informar algú de les mides d'un robot mòbil (fig. 2.a), usant indicadors i colors per informar del parell que està efectuant l'element terminal d'un robot quirúrgic (fig. 2.b), usant un canvi en el color per indicar el volum de la cavitat ventricular en una ecocardiografia (fig. 2.c) o afegint imatges de vehicles virtuals sobre una filmació en un simulador de conducció (fig. 2.d).

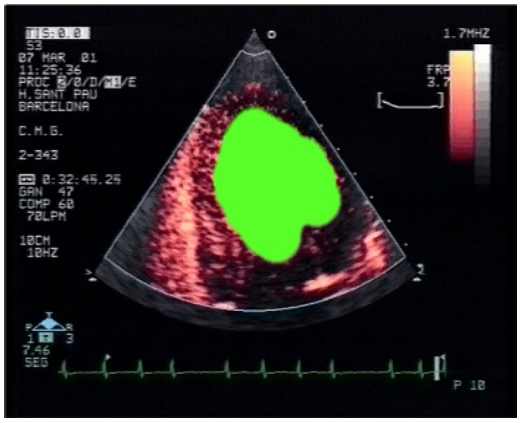
[1] Veure glossari per una definició del terme.



(a) Inserció de mesures de longitud damunt l'estructura d'un objecte



(b) Indicació del parell realitzat per l'element terminal d'un robot



(c) Ressaltat del volum de la cavitat ventricular en una ecocardiografia

(d) Realitat augmentada amb objectes virtuals en escenaris exteriors

Figura 1.2. Diversos exemples de realitat augmentada

En aquesta tesi, es vol explorar una nova forma d'augmentar la realitat amb vistes d'objectes reals. A diferència dels casos anteriors, on la informació a introduir és purament virtual, es volen mostrar dades d'un altre objecte real. Les vistes dels objectes reals seran en principi, imatges obtingudes prèviament, imatges generades a partir d'altres imatges o imatges extretes de models i textures.

En qualsevol cas, l'àmbit d'aquest treball no és el processat d'imatge, la qual cosa ens cenyiria a un món bidimensional, sinó que s'usaran tècniques de visió per ordinador que ens permetran garantir la coherència tridimensional de les vistes dels objectes afegits i de les vistes dels objectes amb l'escena de fons on es fa la realitat augmentada. La generació d'aquestes vistes podrà realitzar-se mitjançant la interpolació entre dues preexistents, la transformació d'una en una altra o l'accés a una base de dades de vistes de l'objecte. En el primer cas, la interpolació es pot fer utilitzant informació de l'estructura tridimensional de l'objecte i garantint que la vista obtinguda per interpolació serà físicament coherent amb la vista real en aquella posició. En el segon cas, la transformació d'una vista en una altra es fa mitjançant tècniques bidimensionals com el *view morphing* que cauen en l'àmbit dels gràfics per ordinador. El tercer cas, d'accés a base de dades de vistes d'un objecte és conceptualment més

senzill però pot esdevenir impossible en funció de la quantitat de vistes a emmagatzemar. En aquest treball s'estudiaran especialment la primera i la tercera i es cercarà un compromís entre qualitat, quantitat de dades a emmagatzemar i temps d'execució.

Per poder fer aquests experiments de realitat augmentada caldran les vistes d'objectes reals i les vistes de l'escena de fons; es necessitarà doncs una certa capacitat d'anàlisi de les imatges obtingudes de l'objecte real i de l'escena on es vol afegir la informació. S'utilitzaran diferents tècniques de visió per ordinador en el mètode desenvolupat, que permetran ordenar les dades, i fer-les coherents amb l'ús de mapes de disparitat, aparellament estèreo i reconstrucció 3D. Així mateix, les imatges font per al mètode tindran uns requeriments en el moment de la captura (control de paràmetres interns i externs de la càmera) que obligaran a usar instruments robotitzats (projector gnomònic, càmeres muntades en robots comercials, etc.). Aquestes dades, juntament amb les d'altres sensors per obtenir la posició i orientació de les càmeres, permetran també fer la correspondència entre el món real i el virtual.

1.1. Definició del marc del treball

Aquest treball cercarà una manera eficient d'introduir en una escena vistes d'objectes que no hi són presents en el moment de la filmació o visualització.

Les escenes on es volen afegir les vistes dels objectes podran ser de qualsevol naturalesa, només caldrà tenir la posició i orientació de la càmera que les està gravant. Si a més a més, es vol que les imatges resultants tinguin certa coherència per a l'observador humà, caldrà que l'escala de l'escena i la dels objectes afegits estiguin en consonància. També serà necessari tenir en compte la disposició d'elements en l'escena per tractar el tema de les oclusions; en els experiments que es plantegen, l'escena serà bàsicament buida i els objectes que hi hagi seran llunyans. En cas contrari, caldria obtenir informació tridimensional precisa de tots els objectes de l'escena, segmentar-los i ordenar correctament els reals i els virtuals en funció de les seves distàncies a la càmera.

En principi, els objectes a tractar seran objectes de dimensions moderades, dels quals se'n pugui obtenir un nombre suficient de vistes (a ser possible des de tots els angles) i que compleixin unes condicions materials com ara la no especularitat, la opacitat i la estaticitat. Si els objectes no són estàtics, de cada una de les vistes caldrà obtenir, no una imatge, sinó una seqüència d'elles. Les concavitats i els forats en els objectes poden ésser un problema i hauran de ser tractats amb cura, com a norma general s'ha d'entendre que si d'una regió d'un objecte, no se n'obté informació aquesta no podrà ser mostrada.

Per al conjunt de dades d'origen, s'haurà d'obtenir unes vistes bàsiques dels objectes amb els que es vol fer realitat augmentada, d'aquestes vistes se n'haurà de tenir una certa informació: altre cop els paràmetres intrínsecs i extrínsecs de la càmera. De les vistes generades, caldrà avaluar la seva qualitat i la seva versemblança. A l'hora de generar les imatges es tindrà un comprimís entre la quantitat d'informació d'origen de l'objecte, el temps de processador necessari per generar una nova vista i la qualitat de la imatge obtinguda. Us sistema que hagi emmagatzemat moltes vistes d'un objecte oferirà major qualitat que un altre. Un sistema on les posicions de les vistes obtingudes compleixin certes regles permetrà una generació més ràpida de les noves vistes.

El sistema d'adquisició de vistes del món on es farà realitat augmentada i el sistema d'adquisició de vistes dels objectes han de tenir certes condicions comunes. Tot

i que en principi el primer serà un sistema obert, que mirarà cap a fora i el segon un sistema tancat que obtindrà vistes al voltant d'un objecte (independentment de que sigui ell o l'objecte qui es mogui), caldrà que la resposta cromàtica dels sensors i les condicions d'il·luminació siguin similars. En cas contrari la composició de les dues imatges seria incoherent. La distància focal, obertura i mida del sensor de la càmera també hauran de ser tinguts en compte per la correcció del factor d'escala. En un cas ideal, el sistema hauria de ser el mateix en ambdós casos.

1.2. Objectius

Aquesta tesi es proposa trobar un mètode per obtenir la millor representació plana d'un objecte, des de qualsevol punt de vista, a partir d'una selecció (si és possible el conjunt mínim necessari) de vistes de l'objecte i informació tridimensional del mateix. Aquesta representació plana, és a dir una imatge, s'utilitzarà per a afegir informació a un escenari, constituint un sistema de realitat augmentada. Aquest objectiu general, serà cobert amb uns objectius parcials que s'enumeren a continuació:

- Es cercarà la manera òptima de sintetitzar noves vistes d'un objecte per al cas proposat, estudiant, comparant i si cal, millorant, els mètodes existents de síntesi o interpolació de vistes.
- Es buscarà un mètode per a trobar el conjunt mínim de vistes que representi suficientment un objecte. Aquest conjunt serà emprat pels mètodes de síntesi esmentats anteriorment.
- S'experimentarà la captura d'aquestes vistes originals i el seu emmagatzemament, per a caracteritzar-ne les problemàtiques específiques que puguin aparèixer.
- Es farà una reconstrucció tridimensional dels objectes, per diferents mètodes, per a complementar el mètode de síntesi de vistes i plantejar-ne d'alternatius.
- S'experimentaran millores, especialment en rendiment i exactitud, en els mètodes de reconstrucció tridimensional d'objectes.
- S'analitzaran els errors de les noves vistes obtingudes, per jutjar la bondat del mètode emprat. Donada la importància d'aquest punt, durant tot el desenvolupament del treball, es guardarà informació per l'avaluació final dels errors. També s'estudiarà per separat cada una de les etapes del mètode (en ordre lògic d'utilització: captura, selecció, emmagatzemament i síntesi) per comparar-los amb altres solucions a les qüestions parcials.
- S'introduiran les noves vistes en sistemes de realitat augmentada per comprovar-ne la utilitat i avaluar-ne el grau de complaença i resposta de l'usuari del sistema. Es treballarà també la correspondència entre el món real i el món virtual, en aquest cas específic en que la informació afegida és una vista d'un altre objecte real no present a l'escena.

1.3. Distribució de Capítols

Aquest treball es troba dividit en deu capítols i el llistat de referències a la bibliografia emprada. A continuació es descriuen de forma breu i individual els continguts de cada un dels deu capítols:

- El primer capítol d'aquesta tesi és la introducció al tema on es vol desenvolupar aquesta tesi i la delimitació del problema a resoldre.

- En el segon capítol es descriuen i discuteixen els treballs d'altres investigadors en realitat augmentada, síntesi de noves vistes i reconstrucció tridimensional d'objectes, veient quines tècniques poden ser útils per a les tasques plantejades.

- En el tercer capítol es parlarà de l'obtenció de la informació fotomètrica que s'utilitzarà per l'experimentació en la resta de la tesi, és a dir, les imatges font amb les vistes dels objectes tractats. Es veurà com aquestes imatges es graven en fitxers, els diferents formats existents pel seu emmagatzemament i com aquest mètode d'organitzar i accedir a les imatges constitueix una primera manera d'obtenir les vistes desitjades.

- En el quart capítol es tractarà la síntesi de noves vistes d'objectes. Primer s'exposaran els conceptes a emprar, s'exposarà el mètode genèric de rectificació de tres vistes i interpolació, i d'aquest mètode, se'n milloraran algunes parts per adequar-lo a la situació plantejada. Aquest serà un altre mètode d'obtenció de vistes: a partir d'algunes preexistents sintetitzar-ne de noves.

- En el capítol cinquè es parlarà d'obtenció d'informació tridimensional d'un objecte a partir de les seves vistes. Com s'haurà vist anteriorment, aquesta informació serà necessària per a la síntesi de noves vistes. Es presentaran els mapes de correspondència i disparitat a partir de l'aparellament estèreo i les seves limitacions. També es mostrarà com obtenir una reconstrucció tridimensional de l'objecte per una via alternativa, per comparar la bondat dels altres mètodes. La representació de les dades tridimensionals amb textura serà una altra via per obtenir les vistes desitjades d'un objecte.

- En el capítol sisè es descriurà la tècnica de *space carving* (procés d'escultura de vòxels) i un mètode per al refinament del volum obtingut per *carving* amb tècniques d'estereovisió. El mètode de *carving* serà descrit completament amb les millores introduïdes en alguns dels seus passos. Es lligarà la implementació del mètode de *carving* amb la selecció de les vistes dels objectes, de manera que es cercarà un subconjunt mínim de vistes per a la síntesi i emmagatzemament de la informació fotomètrica.

- El capítol setè presentarà, a mode de resum de l'exposat, els tres mètodes d'obtenció de vistes identificats, amb les seves característiques i una descripció precisa del seu funcionament.

- En el vuitè capítol es cercaran les fonts dels errors que probablement s'hauran comès, i es cercarà un conjunt de mesures per comparar la bondat de les vistes i la seva fidelitat a l'original.

- En el capítol novè es presentaran els resultats obtinguts: s'avaluaran els mètodes identificats a nivell de temps de procés i dades necessàries i es compararan entre ells amb les mesures d'error trobades al capítol vuitè. També es mostraran dues de les aplicacions que han estat motivació d'aquesta tesi: la realitat augmentada amb vistes d'objectes reals i la presentació remota d'objectes a través d'una xarxa com *internet*.

- El desè capítol mostrarà les conclusions obtingudes en aquesta tesi, les aportacions fetes, les tasques proposades com a treball futur i les línies d'investigació obertes.

2. Estat de l'art

El treball realitzat en aquesta tesi es centrarà principalment en dues àrees: la realitat augmentada, que ha estat una font de motivació pel treball, i la síntesi de vistes d'objectes, íntimament lligada a l'obtenció d'informació tridimensional d'objectes a partir de les seves vistes. La primera àrea es troba a la frontera entre les disciplines de la visió per ordinador i els gràfics per ordinador. La segona es troba clarament dins de l'àmbit de la visió per ordinador.

La realitat augmentada tracta de la barreja d'elements visuals de diferent naturalesa que es presenten a un o més usuaris. Quan predominen els elements d'origen sintètic arriba al llindar de la realitat virtual i els principals motius d'estudi són la generació de gràfics amb ordinador. En el cas d'aquest treball, tots els elements inclosos en la visualització: fons i objectes, seran objectes naturals. Un ordinador els haurà d'enregistrar mitjançant cameres de vídeo, mesurar-ne característiques i finalment, presentar-los a un usuari en una disposició diferent a aquella en que han estat en el món real. Aquestes tasques hauran de ser majoritàriament realitzades amb el coneixement aportat per la visió per ordinador. Anàlogament, en parlar de síntesi de vistes d'objectes, sempre es tractarà de, a partir d'informació tridimensional d'un objecte obtinguda amb tècniques de visió per ordinador, generar altres vistes del mateix.

Els dos temes mencionats seran tractats segons diferents punts de vista: mentre la realitat augmentada serà utilitzada com a marc d'aplicació per al treball realitzat, la síntesi de vistes d'objectes comprendrà el cos principal d'aquesta tesi, amb la reconstrucció tridimensional d'objectes que serà emprada com una eina per a la millora de la síntesi de vistes. En aquest estat de l'art, es revisaran els dos temes cercant els treballs, autors i resultats més propers a aquest enfocament, també es donarà al final una pinzellada de l'estat de l'art en obtenció de la informació tridimensional d'un objecte a partir de les seves vistes.

2.1 Realitat augmentada

Com ja s'ha dit, el terme realitat augmentada es refereix a la combinació d'informació visual del món real amb la generada mitjançant ordinadors. Aquesta definició genèrica, proposada per Wellner [Wellner 93], estava lligada a l'ús d'un HMD i sorgia com alternativa a la realitat virtual. La definició ha estat ampliada posteriorment a l'afegit d'informació visual a qualsevol entorn real mitjançant gràfics per ordinador

[Milgram 94] i encara més ampliada a qualsevol sistema interactiu que combini el que és real amb el que és virtual [Azuma 97] sempre i quant estigui en correspondència.

L'increment del nombre d'aplicacions on es combina el que és real amb el que és virtual, va fer aparèixer un concepte més ampli anomenat "*mixed reality*" [Milgram 99] en el que es defineix tota una gradació de situacions des del món real fins al món virtual (veure figura 2.1).

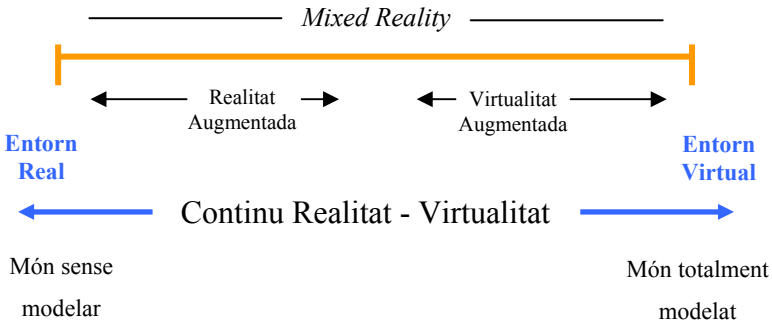


Figura 2.1 Definició del concepte *Mixed Reality* que engloba tot l'espectre entre el món real i el món virtual.

Dins d'aquest espectre, es focalitzarà l'estudi en les aplicacions de realitat augmentada on, en un entorn majoritàriament real, s'hi introdueix alguna informació virtual. Cal ara doncs, parlar de dos conceptes: com es fa la barreja d'informació real i virtual, i com s'alineen aquestes informacions perquè l'usuari les percebi correctament.

La presentació de la informació a l'usuari es planteja habitualment amb l'ús d'un casc de realitat virtual del tipus HMD o unes *see-through glasses*. El primer dispositiu consisteix en unes pantalles situades davant dels ulls de l'usuari on s'hi dibuixa l'escena augmentada (veure aplicacions industrials en el sector de l'aviació [Boeing 97] o d'un consorci empresarial alemany [Arvika 03] a la figura 2.2). El segon té un vidre que deixa passar la imatge del món real i damunt d'aquest vidre s'hi projecta la informació afegida. Un report complert de dispositius per realitat augmentada es pot trobar al report elaborat per Fuchs i Ackermann [Fuchs 99].

Figura 2.2 Exemples d'utilització de dispositius de realitat augmentada en el món industrial (consorci industrial Arvika, companyia Boeing). Les aplicacions, molt similars, consisteixen en mostrar a un operador informació addicional dels dispositius que manipulant.

D'altra banda, l'alineament entre el món real i el món virtual, és el que s'anomena posada en correspondència o en terminologia anglesa *registration*. Consisteix bàsicament en fer que l'estructura tridimensional del món virtual correspongui amb la del món real perquè les informacions que s'aporten es representin en el lloc correcte, i també és necessari conèixer la posició de l'observador i cap a on

està mirant per saber com projectar el món virtual. Si l'usuari du al damunt un casc de realitat virtual un sensor de posició i orientació de sis graus de llibertat [Polhemus 05] pot resoldre aquesta necessitat, si hi du una càmera es poden emprar tècniques de visió per ordinador per resoldre la posició i orientació a partir de cert nombre de punts coneguts [Fischler 81], amb un mínim de tres [Gao 03]. Una altra opció és la de situar marques damunt l'estructura del casc o la persona i amb una càmera externa calcular la seva posició i orientació.

Les temàtiques dels treballs en realitat augmentada versen des dels més propers a la indústria com els mostrats, fins als més propers a l'entreteniment com jocs amb realitat augmentada [ARQuake 06] [Piekarski 02] o pel·lícules amb ús de realitat augmentada pels efectes especials [BBC 00] (veure figura 2.3).

Figura 2.3 Aplicacions de la realitat augmentada a l'entreteniment, esquerra (jugador) i centre (vista subjectiva): joc ARQuake desenvolupat per la Universitat del Sud d'Austràlia. A la dreta fotograma de la pel·lícula "*walking with dinosaurs*" produïda per la BBC.

Quan es treballa en entorns exteriors, es solen emprar models CAD de l'escena per ajudar a la posta en correspondència, això és habitual en treballs relacionats amb l'arquitectura com l'estudi d'un pont mostrat a l'esquerra de la figura 2.4, o en simuladors d'escenes de trànsit amb realitat augmentada com el mostrat també a la dreta de la figura 2.4 [Arboleda 02]. En referència al sistema de coordenades del model CAD es solen situar els objectes a sobreposar a les escenes reals.

Figura 2.4 Exemples de realitat augmentada en entorns exteriors, amb correspondència amb models CAD: Pont sobre el Tàmesi (estudi de Norman Foster), i tramvia virtual sobre la avinguda Diagonal de Barcelona (departament ESAII, UPC).

La popularització de les aplicacions amb realitat augmentada, ha fet que apareguin llibreries de codi obert que inclouen funcions de captura d'imatges, posada en correspondència, detecció de marques i seguiment de marques per visió, superposició

d'imatges, etc. com ARToolkit [ARToolkit 06], que permeten desenvolupar en molt poc temps programes domèstics o educatius de realitat augmentada (veure figura 2.5).

Figura 2.5 Exemple d'aplicació desenvolupada amb la llibreria ARToolkit.

Fins ara s'han mostrat aplicacions on els objectes superposats interactivament són clarament elements virtuals com informació alfanumèrica, dibuixos CAD, o figures per jocs. Sols en les escenes de la pel·lícula de la figura 2.3 s'han presentat objectes fotorealístics, però el seu temps de projecció és encara molt gran i dista de poder constituir una aplicació interactiva.

Els darrers anys però han començat a sortir empreses que exploten comercialment les possibilitats de la realitat augmentada. Concretament, l'empresa Total Immersion [Total Imm. 06] ofereix aplicacions de realitat augmentada a temps real, per programes de televisió, anuncis comercials, transmissions esportives, etc. La figura 2.6 mostra imatges de dues aplicacions desenvolupades amb realitat augmentada per aquesta empresa per a televisió.

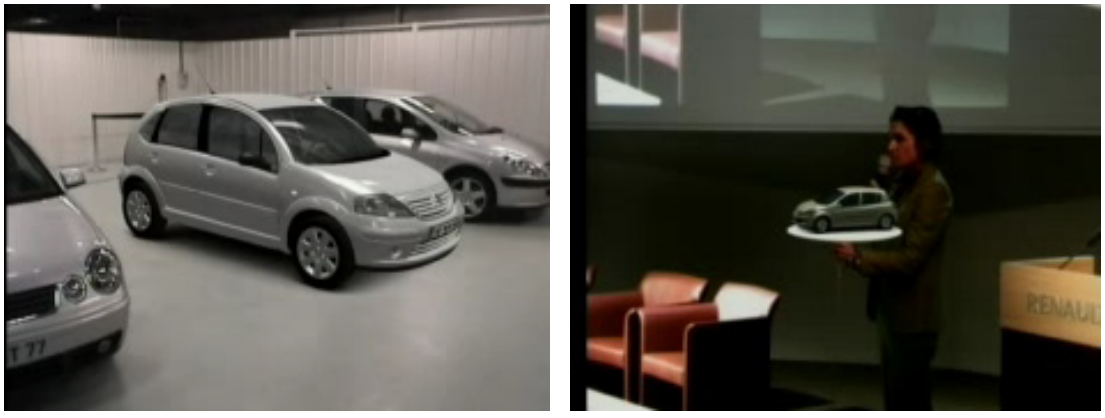


Figura 2.6 Exemple d'aplicacions de la realitat augmentada al mercat de la televisió i publicitat.

Els darrers treballs presentats, aprofitant la contínua millora de les prestacions del maquinari dels ordinadors, introdueixen objectes cada cop més fotorealístics en la realitat augmentada. Quan aquests objectes no tenen l'origen en models CAD, sinó en imatges tretes d'objectes reals, caldrà plantejar-se quines tècniques hi ha per obtenir la

informació, extreure'n (si s'escau) l'estructura tridimensional i generar noves vistes de l'objecte a partir d'aquestes dades.

2.2 Síntesi de vistes. Utilització del coneixement de l'estructura tridimensional de l'objecte

La generació de noves vistes d'una escena a partir d'un conjunt d'imatges (en general dues o més) que s'han obtingut o s'estan obtenint de la mateixa escena, s'anomena síntesi de vistes [Seitz-Dyer 95]. En referir-se a síntesi de vistes d'objectes, sempre es parlarà d'objectes reals, existents en el món físic, dels quals s'ha obtingut informació fotomètrica mitjançant una o més cameres, i dels quals es volen obtenir noves vistes des de llocs on no s'ha ubicat una camera o element captador. En el cas particular en que el nou punt de vista hagi de trobar-se en la línia recta que uneix dues vistes de referència, es parlarà d'interpolació de vistes. La interpolació de vistes, presentada per Chen i Williams el 1993 [Chen 93] és doncs, més restrictiva i es pot considerar com un cas particular de la síntesi de vistes.

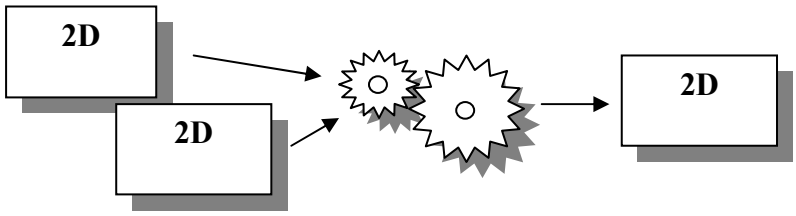
Cal distingir dues aproximacions a la síntesi: una en la que s'utilitza informació tridimensional de l'objecte en qüestió, obtinguda, per exemple, mitjançant l'aplicació de tècniques de visió per ordinador, i una segona aproximació en la que s'empra només la informació bidimensional de les imatges. Mentre la primera garanteix la consistència física de les vistes obtingudes [Scharstein 99], la segona tècnica no garanteix que les imatges sintètiques siguin físicament coherents amb la vista de l'objecte real en aquella posició sinó es compleixen unes condicions de monotonicitat en les imatges i no especularitat dels objectes [Seitz-Dyer 95],.

En la utilització de la informació tridimensional obtinguda de l'objecte per a la síntesi de vistes, cas que garanteix la coherència amb el món físic de les imatges sintetitzades, i en el que es focalitzarà l'estudi, es troben dues tendències prou diferenciades: autors que expliciten la informació 3D amb la construcció dels models i les estructures de dades necessàries, i autors que no expliciten aquesta informació, però que l'usen implícitament per als càlculs de generació de la nova imatge. Un cop obtingut el model tridimensional dels objectes, també es podrà conduir aquesta informació cap a una eina de CAD per incloure-la en aplicacions de realitat augmentada com les mostrades a l'apartat 2.1.

En base a aquestes distincions fetes es repassaran tot seguit els treballs més destacats en l'àmbit de la síntesi de vistes segons tres categories: mètodes de síntesi de vistes sense utilització de la informació tridimensional, mètodes de síntesi de vistes amb explicitació de la informació tridimensional i finalment, mètodes que empren aquesta informació de manera implícita però no arriben a representar-la. En els mètodes que empren informació tridimensional, es mostra com els autors obtenen aquestes dades.

Finalment es mostrarà també un resum de les tècniques més emprades per l'obtenció de l'estructura tridimensional d'una escena a partir de les dades capturades per una o més càmeres.

2.2.1 Síntesi sense utilitzar la informació 3D



Algoritmes com *view morphing* [Seitz 99] o els coneguts programes de *morphing* de cares (veure figura 2.7), són capaços de generar noves vistes d'objectes només a partir d'un conjunt d'imatges dels mateixos. Tan sols requereixen una aplicació que relacioni els píxels d'una imatge original amb els d'una altra (en general es tractarà d'una aplicació bijectiva). En base al mètode gràfic de *morphing* (en anglès, canvi de forma) entre dues imatges, Steven Seitz i Charles Dyer de la Universitat de Wisconsin, s'han plantejat demostrar que les imatges generades per interpolació 2D entre dues vistes corresponen també a vistes físicament vàlides d'alguns dels punts intermitjos. Partint dels treballs de Chen i Williams en que demostraven que la interpolació de vistes a partir d'imatges seria més ràpida que el *render* (en anglès, dibuix per projecció) a partir de models 3D [Chen-Williams 93] [Seitz 97], els autors demostren que sota unes condicions de monotonicitat en les imatges i de no existència de superfícies especulars en les mateixes (aplicació del model Lambertià), "la interpolació entre imatges bidimensionals d'un objecte és un mecanisme vàlid d'interpolació de vistes" [Seitz-Dyer 95].

Figura 2.7 Exemple d'aplicació de la tècnica de *morphing* entre les cares de dos individus (imatges dels extrems).

D'aquesta manera, a partir d'un conjunt originari d'imatges, i amb les operacions de rotació, translació i *warping* (en anglès, deformació) d'imatges, es podrà aplicar l'algoritme de *view morphing* per a la obtenció de vistes sintètiques. Un altre avantatge defensat pels autors [Seitz 97] és l'estalvi d'espai a disc i temps de processador a l'hora d'emmagatzemar les dades respecte als mètodes que extreuen informació tridimensional més enllà del món de les imatges. La interpolació de vistes queda doncs tancada en el domini de les imatges, són transformacions entre matrius bidimensionals de punts.

Aquesta tècnica de *view morphing* tan pot ser usada per imatges calibrades com no calibrades [Seitz-Dyer 96], en el cas de tenir un escenari monòton i Lambertià. La condició de monotonia imposa que si donats dos píxels en una imatge, un és a l'esquerra de l'altre, en una altra vista de la mateixa escena, es mantindrà aquesta

relació [Seitz 97]. El model Lambertià imposa que la reflexió dels raigs de llum en qualsevol de les superfícies existents a l'escena serà uniforme en totes les direccions; contràriament al model especular en que la llum arriba a una superfície de tipus mirall i es reflexa en una única direcció.

En el cas de tenir un model de les càmeres perfectament calibrat, amb unes relacions bàsiques de geometria es troba la correspondència entre diferents regions de la imatge. Pel cas de no tenir calibrades les cameres, l'autor deixà un conjunt de factors com a paràmetres en les equacions de les relacions geomètriques i els resol trobant posteriorment correspondències entre punts. Aquesta correspondència entre els píxels la resol o bé a mà, amb interacció de l'usuari, o extraient característiques amb processat d'alt nivell que permetin relacionar algunes característiques de les diferents imatges. La figura 2.8 mostra el resultat de la tècnica de *view morphing* de la Universitat de Wisconsin [Seitz 99] aplicada a dues vistes d'una capsa de tiretes de les quals se n'ha sintetitzat una tercera corresponent a un punt de vista intermig.

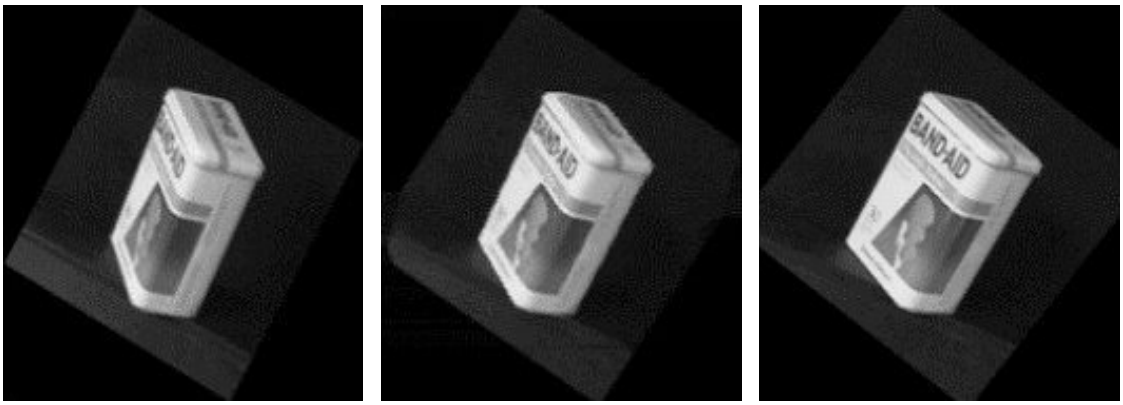
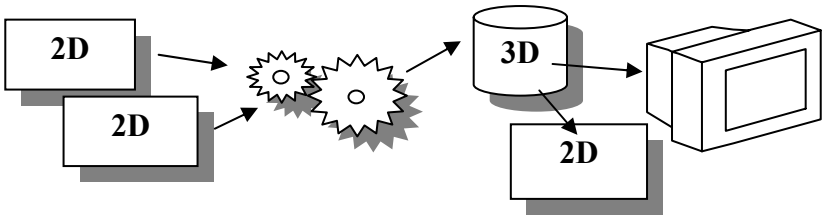


Figura 2.8 Exemple d'aplicació de la tècnica de *view morphing* entre dues vistes d'un objecte de geometria regular (la vista central està interpolada just al mig de les dues laterals).

L'algoritme de *view morphing* requereix doncs, en general, [Beier-Neely 92] tenir una correspondència entre tots els píxels de la imatge originària i els de la imatge destí; Seitz i Dyer han demostrat que la correspondència entre uns conjunts de píxels especials, és suficient per la generació de la transformació. Aquests píxels són els dels contorns de les regions amb característiques similars. Assignada la correspondència entre els contorns en les dues imatges, es demostra pràcticament [Seitz 97] que els píxels interiors segueixen força correctament el procés de transformació.

De tota manera, les restriccions en la naturalesa i geometria de l'objecte imposada per l'aplicació del mètode i les dificultats que poden aparèixer en el càlcul de l'aplicació entre els punts d'una vista i els de l'altra, limiten l'ús de l'algoritme de *view morphing* a uns escenaris concrets. Quan es volen superar algunes d'aquestes restriccions, fent que els punts siguin aparellats segons un criteri de correspondència estereoscòpic ja s'abandona el món de les dues dimensions per emprar, de forma implícita, la informació de l'estructura tridimensional de l'escena.

2.2.2 Síntesi amb explicitació de la informació 3D



A continuació es mostraran els treballs dels autors que empren les imatges donades per una, dues o més càmeres per extreure la informació tridimensional de l'escena i a partir d'aquesta sintetitzar qualsevol nova vista de l'escena. Com es veurà es tracta de, en general, algoritmes costosos que requeriran una gran capacitat de procés per arribar a generar aplicacions interactives.

2.2.2.1 Eixam de cameres: *Virtualized reality*

El grup de recerca en robòtica de la universitat de Carnegie-Mellon a Pittsburgh, en col·laboració amb el centre de recerca en Intel·ligència Artificial de Bangalore (Índia) ha estat desenvolupant un equipament per a la construcció d'imatges sintètiques a partir d'un model tridimensional del món i les textures obtingudes per les mateixes imatges [Narayanan-Kanade 97]. L'equipament consisteix en la utilització d'un gran nombre de càmeres observant la mateixa escena des de diversos punts de vista, càmeres connectades a una xarxa de processadors.

L'ajust entre les imatges reals i el model virtual és anomenat el problema de *registration* [Azuma 99] (explicat a l'apartat 2.1 d'aquest capítol). Per millorar aquest encaix, els investigadors han estudiat la construcció de móns virtuals a partir de les dades del món real, utilitzant tècniques de *dense stereomatching* [Narayanan-Kanade 98], consistents en la creació de mapes de correspondència entre molts punts de les diverses càmeres emprades (l'obtenció de mapes de disparitat densos entre imatges ha estat una font permanent de treball pel grup de Kanade, arribant a dissenyar fins i tot processadors específics per a la tasca [Kanade 96]).. Amb aquesta informació i els paràmetres de calibració de les càmeres es podrà construir el model tridimensional de l'escena.

Els resultats obtinguts són satisfactoris (veure figura 2.9), malgrat que segueix existint un problema greu de "costures" amb els elements del món real. Aquest problema de les "costures" comprèn tots els camps on es barregen elements reals i virtuals, des de l'*augmented reality* (realitat augmentada) fins a l'*augmented virtuality*, passant per totes les etapes intermèdies segons la taxonomia de Milgram i és un dels principals cavalls de batalla en el camp de la *Mixed Reality* (barreja de realitat i virtualitat) [Ohta-Tamura 99]. La tendència actual per la superació de les costures és l'aplicació d'una certa difuminació als colors dels contorns dels objectes utilitzats.

Les principals aplicacions donades fins al moment pel grup de Kanade als treballs en *virtualized reality* (TM) [Kanade 99] han estat en el camp recreatiu, amb convenis amb empreses de software (EA) i cadenes de televisió per la creació de videojocs i millora dels efectes especials introduïts en les retransmissions esportives.

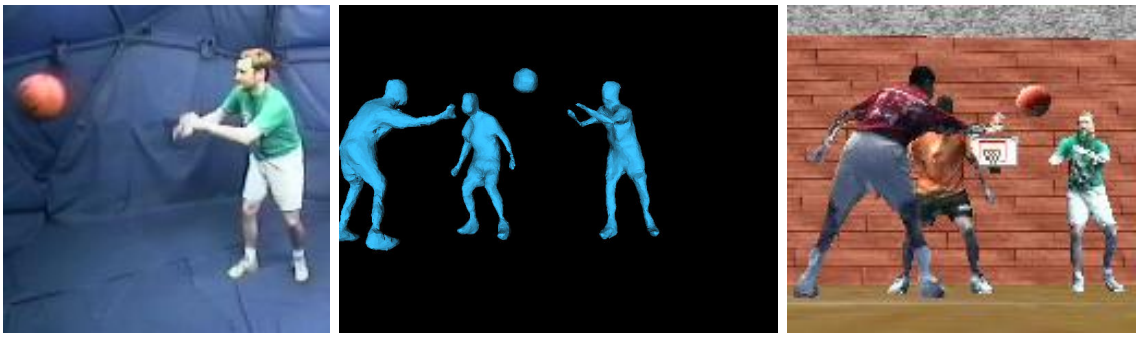


Figura 2.9 Exemple d'aplicació de *virtualized reality*: a l'esquerra es captura amb diverses càmeres el moviment d'una persona, de la que es reconstrueix a temps real l'estructura tridimensional. Aquesta estructura s'introdueix en un model virtual (centre) al que s'apliquen textures per la visualització final (dreta).

2.2.2.2 Eixam de cameres: teleconferències

En un treball conjunt de les universitats de Carnegie-Mellon (grup de Takeo Kanade), Carolina del Nord (Fuchs, Bishop, Arthur i McMillan) i de Pensilvània (Bajcsy, Lee i Farid) es proposa altre cop l'ús d'un nombre gran de cameres (mar de cameres segons els autors) per obtenir (s'intenta que a temps real) la informació fotomètrica i de profunditat d'un escenari [Kanade-Fuchs 94]. Aquesta informació, juntament amb un model prèviament construït de l'entorn, serà enviada a un usuari remot, el qual podrà reconstruir qualsevol punt de vista en l'escenari mitjançant tècniques de reprojecció i aplicació de textures. Aquesta, pretenen els autors, serà la base de les teleconferències en el futur, amb aplicacions en el camp de la robòtica, la medicina i el mercat de la televisió (veure figura 2.10).

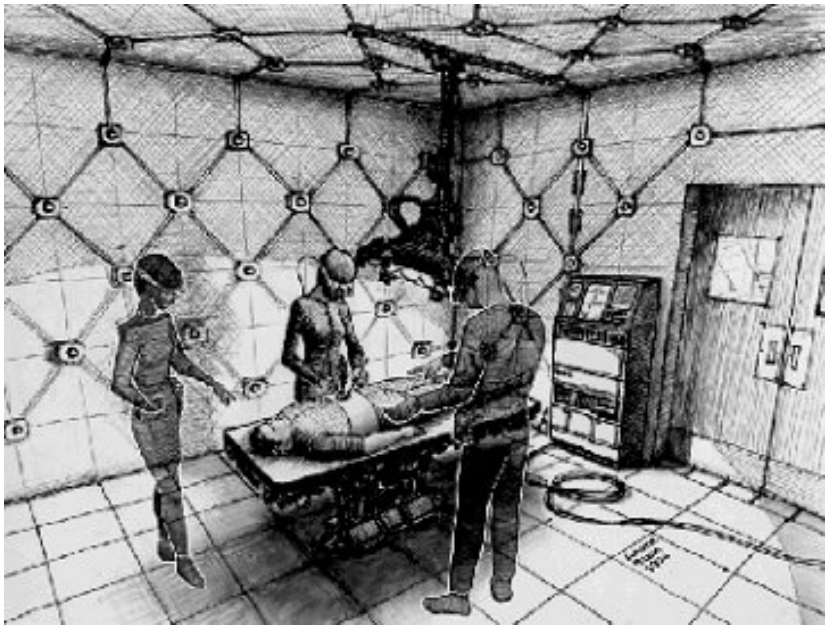


Figura 2.10 Representació de la utilització de l'anomenat "mar de càmeres" per la realització de teleconferències. La reconstrucció de l'entorn tridimensional permetria que el receptor pogués moure lliurement el seu punt de vista en l'escena, cosa especialment interessant en el teleensenyament de, per exemple, medicina.

L'interès comercial per l'evolució de la televisió en el futur, amb pretesa capacitat de projecció tridimensional [AS 99], ha generat la dedicació d'esforços per part dels departaments de recerca d'empreses: *Sony Corporation*, *Canon Inc.*, *Actuality Systems*, i d'universitats en la investigació de la síntesi d'imatges [Fuiji-Harashima 94]. Això és degut a que es vol saber quin és el conjunt mínim de vistes que caldrà capturar per a poder reconstruir-ne qualsevol a voluntat del receptor, amb interès en minimitzar el volum de transmissió de dades [Naemura-Harashima 99] [Seitz 97]. Respecte als sistemes de representació, per a aquest nou model de televisió (entenguis també per teleconferències) es proposen dispositius HMD, ulleres estereoscòpiques basades en vidres polaritzats, projectors hologràfics, etc. [Fuchs-Ackerman 99] [AS 99]

2.2.2.3 Reconstrucció tridimensional amb models d'equacions diferencials parametritzades (EDP)

Aquest treball és el fruit del conjunt de treballs desenvolupats a l'INRIA (acrònim de l'Institute National de Recherche en Informatique et Automatique) a França., i el grup d'Olivier Faugeras (traslladat de França al MIT de Boston) en que s'obté la reconstrucció de l'estructura tridimensional d'una escena a partir de les imatges capturades i després es sintetitzen noves imatges a partir de la reprojecció d'aquesta informació. La figura 2.11 mostra els resultats de la recuperació d'estructures en escenes urbanes [Faugeras-Laveau-Robert 95]. Les imatges de les escenes urbanes són capturades amb cameres no calibrades i es plantegen sistemes d'equacions diferencials parametritzades per reconstruir la geometria epipolar de l'escena. Les incògnites d'aquestes equacions es van resolent mitjançant dades que aporta el *matching* de punts de l'escena fet per diferents tècniques de visió per computador. Finalment es troben per cada camera les matrius de projecció perspectiva i es dedueix l'estructura geomètrica dels elements de l'escena.

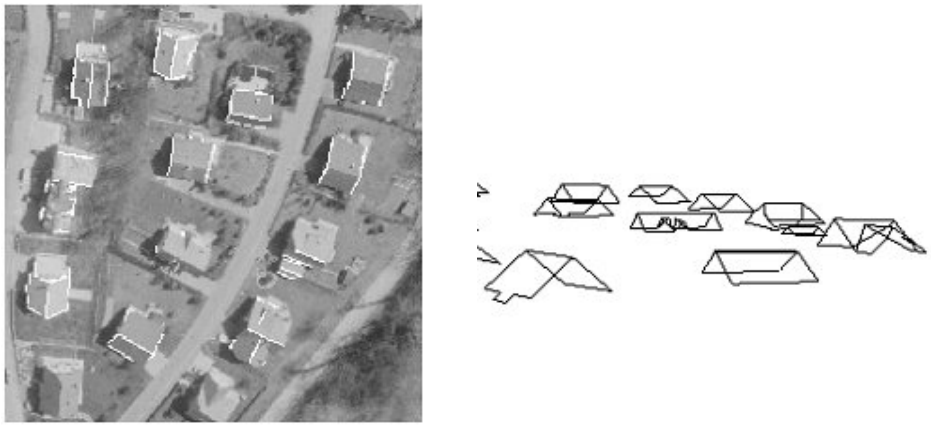


Figura 2.11 Reconstrucció de part de l'estructura tridimensional d'una escena urbana a partir de punts identificats per diferents càmeres (ressaltats a la vista de l'esquerra) i el model d'equacions diferencials parametritzades.

Un altre treball desenvolupat pels grups de recerca a l'INRIA mitjançant els sistemes d'equacions diferencials parametritzades (EDP) és la recuperació de l'estructura tridimensional (en sentit estricte, les anomenades dues dimensions i mitja) d'una cara de la qual s'han pres diferents imatges (en general 2 o més). Primer es planteja l'equació d'una superfície que "flota" a l'espai, i que es va ajustant

iterativament a l'estructura de l'escena capturada. Donat que les imatges originals s'imposen com a condicions al sistema, aquest tendeix a minimitzar l'error entre la projecció 2D de la superfície que s'està modelant i les imatges preses com a patró. Aquest sistema té una tolerància intrínseca a fallades de calibració del sistema, i per tant les cameres amb que s'han adquirit les dades no han de ser calibrades, sols han de tenir uns paràmetres semblants [Faugeras-Keriven 96]. La figura 2.12 mostra el resultat de la convergència de la superfície sobre un rostre del qual s'han pres dues fotografies [Faugeras 99]. No cal dir, que aquest mètode, tot i donar uns resultats espectaculars, té un cost de computació altíssim i de moment no es pot pensar en aplicacions a temps real basades en ell.

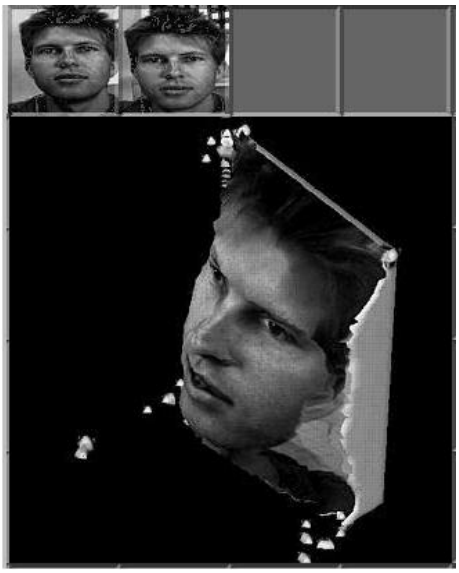


Figura 2.12 Reconstrucció de part de l'estructura tridimensional d'una cara mitjançant el mètode de minimització d'errors en una superfície parametritzada amb EDP.

2.2.2.4 *Generalized Voxel Coloring* (GVC) i *Space Carving*

Un dels mètodes clàssics per a la reconstrucció tridimensional d'objectes o escenes consisteix en crear un espai tridimensional de vòxels i determinar per a cada un d'ells si és o no visible per les càmeres i quin ha de ser el seu color.

Les dues aproximacions més emprades pel problema són els mètodes de *Voxel Coloring* (traduïble per acoloriment de vòxels i exemplificat a l'article [Seitz-Dyer 97]) i el de *Space Carving* (traduïble com a esculpit de l'espai, i tractat a l'article [Kutulakos-Seitz 98]).

El primer mètode exigeix unes restriccions en la ubicació de les cameres per simplificar la computació de la visibilitat; concretament, les cameres solen estar col·locades sempre a un costat de la imatge. El segon, el de *space carving*, permet generalitzar la ubicació de les cameres, però presenta problemes a l'hora de determinar la consistència del color dels *voxels*. De la mateixa manera que el primer mètode, el procés consumeix molt de temps de còmput.

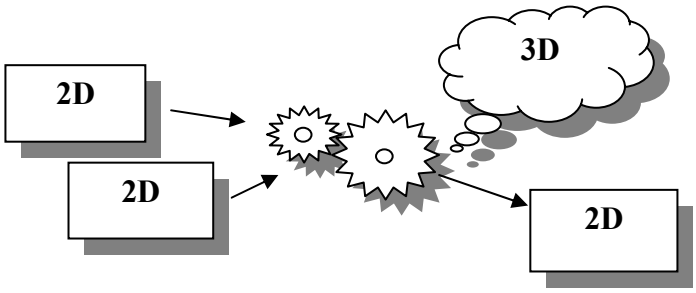
Recentment, s'ha desenvolupat una generalització de l'algoritme d'acoloriment de vòxels que permet qualsevol ubicació del conjunt de cameres [Culbertson-Malzbender 99]: és l'anomenat *generalized voxel coloring*. Aquest algoritme requerirà tenir un conjunt de cameres calibrades per determinar la projecció de les imatges, i per a

cada vòxel crea una estructura de dades que avalua la seva visibilitat en funció de l'estat dels vòxels veïns. També avalua per a cada vòxel la consistència del color que li és assignat: si a un mateix element se li intenten atorgar colors massa diferents aquest *voxel* serà eliminat de la estructura tridimensional de l'escena.

La figura 2.13 mostra un conjunt de quatre imatges naturals preses pels autors i la seva reconstrucció per l'algoritme de *space carving* i de *generalized voxel coloring* respectivament. Tot i que el mètode de GVC obté uns resultats molt millors que la formulació clàssica dels mètodes de *space carving* o *voxel coloring*, té un temps de còmput de l'ordre de desenes de minuts, variant en funció del volum a reconstruir.

Figura 2.13 Reconstrucció de l'estructura tridimensional d'una escena a partir de quatre vistes de la mateixa (grup superior de vistes), pel mètode de *space carving* (esquerra) i *generalized voxel coloring* (dreta).

2.2.3 Síntesi sense ús explícit de la informació 3D



2.2.3.1 Projecció de plans a temps real

L'objectiu principal d'aquest mètode presentat recentment per Yang, Welch i Bishop de la universitat de Carolina del Nord (UNC) és aconseguir la generació de noves vistes a *video-rate*. L'algoritme plantejat [Yang-Welch-Bishop 02] consisteix en l'adquisició simultània de diversos punts de vista d'una escena des de vàries càmeres connectades a diferents ordinadors.

Seguidament es determina el punt de vista des d'on es vol sintetitzar la nova imatge, i es crea un conjunt de plans paral·lels al pla de la CCD virtual. Mitjançant targes gràfiques convencionals (Nvidia GeForce [Nvidia 06]) es determina la projecció de cada una de les imatges en tots els plans i llavors, per a cada píxel, es mira la distribució de colors projectats (mitjana i variança). Aquesta informació s'utilitza per a calcular si un píxel ha de ser o no pintat en la imatge resultant (si la variança és massa alta se'l considera inconsistent, a l'estil del mètode de *voxel coloring*) i quin color ha de tenir aquest píxel.

La figura 2.14 (pàgina següent) mostra el sistema d'adquisició emprat a l'experiment, les imatges preses i el resultat de la reprojecció de la nova imatge. Tot i la bona resposta en temps del sistema, no garanteix la correctesa física de les imatges generades i es troba limitada tant la ubicació de les cameres per la captura com l'espai on ubicar els nous punts de vista.

2.2.3.2 Models geomètrics projectius orientats

Com ja s'ha vist a l'apartat 2.2.2, en l'àmbit de la síntesi d'imatges a partir de vistes prèvies d'una escena, un dels treballs amb resultats més espectaculars i amb una base matemàtica més sòlida és el realitzat a l'INRIA. Concretament, a les seus de Rhône-Alpes i Sophia-Antipolis, iniciadors dels estudis teòrics, on en el cèlebre article de 1993: "*What can two images tell us about a third one?*" [Faugeras-Robert 93], es posaven les bases pel treball en interpolació d'imatges en base a les propietats de la geometria epipolar.

Com a exemple, es mostra la figura 2.15 [Laveau 99] on a partir de tres vistes d'una escena de carretera i amb un temps de còmput (SGI Indy, any 2000) de mig segon per imatge, es genera una vista interpolada en color de l'escena. El mètode emprat en aquest experiment, utilitza per a la representació de les dades unes estructures especials que codifiquen models geomètrics projectius orientats [Laveau-Faugeras 97]. Aquests permeten trobar les matrius fonamentals de cada vista i així optimitzar la síntesi de noves vistes amb les propietats de la geometria epipolar, en que es preserva l'estructura tridimensional de l'escena. El mencionat temps de procés per a la generació de cada

nova vista momentàniament impedeix la implementació del mètode per a representacions a *video-rate*, però amb futures optimitzacions i especialitzacions del hardware es podria arribar a assolir-ne els requeriments.

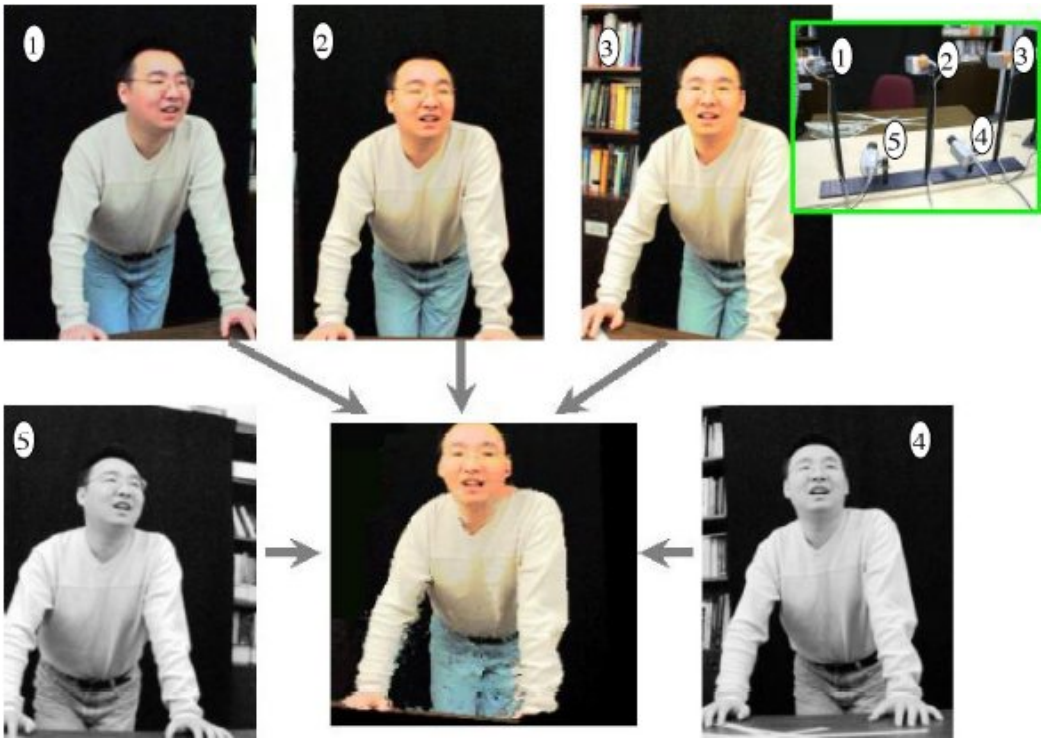


Figura 2.14 Síntesi d'una vista sintètica amb el mètode de projecció de plans a temps real. A partir de 5 càmeres (dalt, dreta), es genera a *video-rate* la vista sintètica (baix, centre).

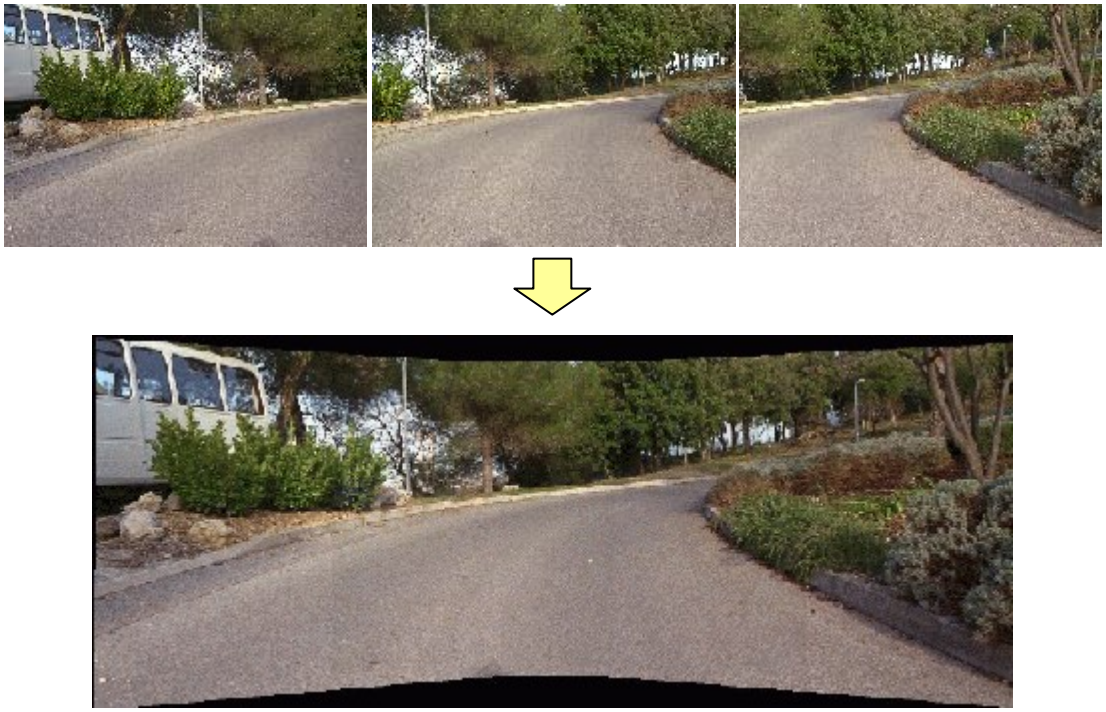


Figura 2.15 Síntesi d'una nova vista a partir de tres originals i models projectius orientats.

2.2.3.3 Joint View Triangulation (JVT)

També a l'INRIA es troba un altre grup de treball que sintetitza noves vistes d'escenes sense necessitat de reconstruir l'estructura 3D de l'escena. En aquest cas l'element clau és també la forma de representar les dades capturades de les imatges originals [Laveau-Faugueras 94]: un cop capturades les imatges, es generen mapes de correspondència mitjançant algoritmes de visió de computador d'alt nivell, i seguidament la informació obtinguda es guarda de manera òptima per facilitar la síntesi de noves vistes.

En aquest cas [Lhuillier 99], es troben les correspondències de grups de píxels amb ús de l'algoritme de *region growing* (creixement de regions), i es guarden aquestes dades en mapes JVT (*Joint View Triangulation*). Aquests, sobre la informació dels objectes segmentats, guarden les seves posicions relatives i la correspondència amb les regions segmentades de l'altra (o altres) imatges. En la figura 2.16 es mostren les imatges originals d'una escena natural, els píxels marcats de les regions que fan matching amb píxels de l'altre imatge, l'estructura usada pel *region growing*, els dos mapes JVT de les dues imatges i finalment tres noves vistes de l'escena generades per reprojecció dels mapes JVT.

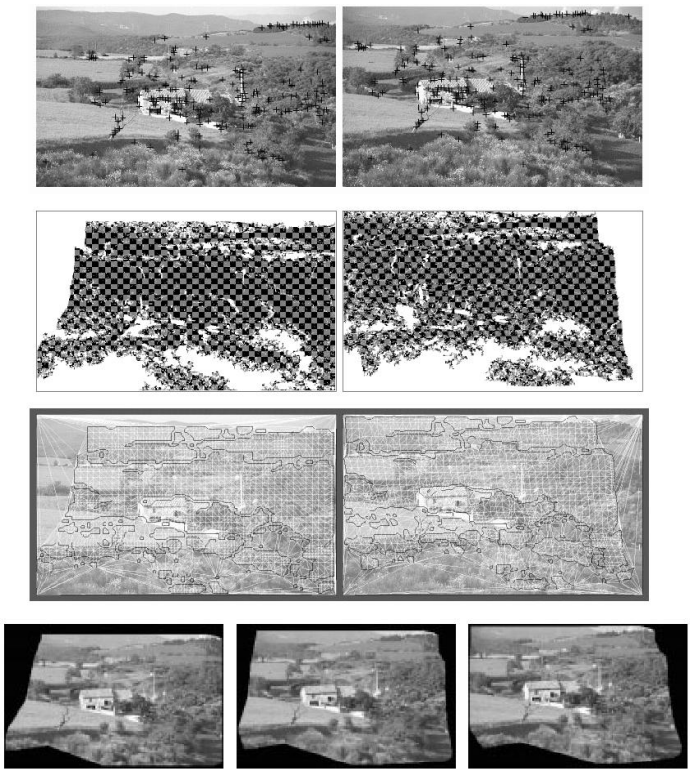


Figura 2.16 Síntesi de tres noves vistes d'una escena (part inferior) a partir de dues originals (part superior) i el mètode de Joint View Triangulation.

2.2.3.4 Mapes panoràmics

Entre els grups que treballen sense model previ de l'escenari a tractar, però que realitzen un processat d'alt nivell de la imatge per a la síntesi de noves vistes, trobem els investigadors de les Universitats de Tokio i North Carolina que, amb les imatges

adquirides, generen uns mapes panoràmics amb geometria cilíndrica. En el cas de la universitat de Tokio [Hirose 99] centenars de mapes panoràmics (veure figura 2.17, part superior) d'una ciutat són capturats des d'un vehicle on s'han muntat un conjunt de vuit cameres (veure figura 2.17, part inferior).

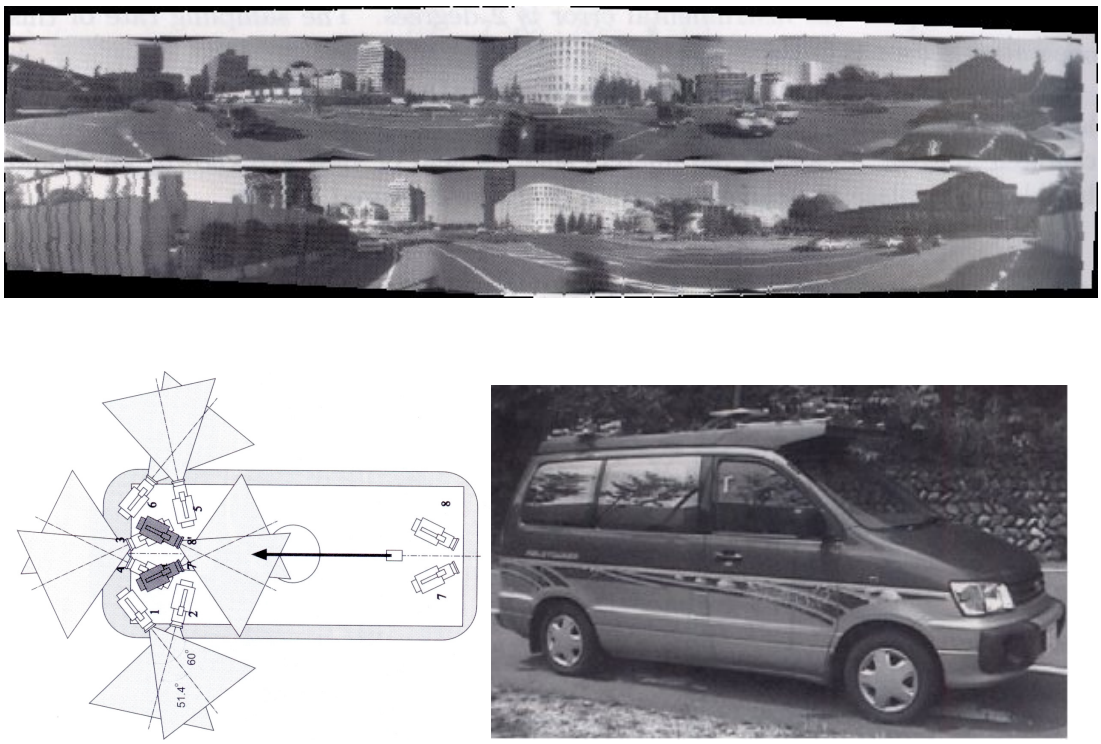


Figura 2.17 Mapes panoràmics en geometria cilíndrica obtinguts d'una ciutat (part superior), amb un vehicle equipat amb vuit càmeres (part inferior de la figura).

El procés de "cosit" de les imatges per a la generació dels mapes panoràmics es fa automàticament per detecció de característiques de les imatges, que tenen regions encavalcades. A partir d'aquestes imatges s'interpolen de forma continua noves vistes, donant la sensació que s'està navegant per la ciutat. Aquesta tècnica s'anomena *walk through* i gràcies a la geometria cilíndrica, a part de navegar per la ciutat es pot girar amb llibertat l'angle de visió en el pla horitzontal, tècnica anomenada *look around*.

La combinació d'aquestes dues tècniques dóna una gran sensació de realisme a l'usuari del sistema, però té un cost computacional molt gran i necessita una altíssima capacitat d'emmagatzemament. En el cas de la universitat de Tokio les dades són gravades i llegides des d'un làser disc. Per a la ubicació en l'espai del vehicle, el sistema compta amb receptors GPS, sensors de magnetisme terrestre i sensors d'inclinació. La síntesi de noves vistes es realitza mitjançant un algorisme de matching d'objectes entre les diferents vistes filmades. Un cop realitzat el matching (figura 2.18) l'autor presumeix que els objectes detectats es troben en plans a l'espai [Hirose 99] i interpola les noves vistes mitjançant reprojecció d'aquests objectes sobre el fons de l'escenari. Les imatges ja sintetitzades són projectades a l'usuari en tres pantalles gegants donant-li un angle de visió de 180º. L'usuari controla la navegació mitjançant un dispositiu *joystick* (veure figura 2.19).

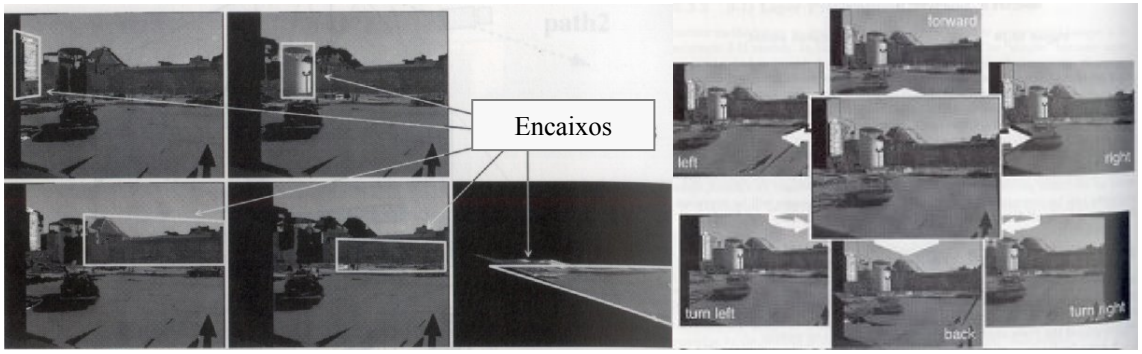


Figura 2.18 Procés d'aparellament de regions i interpolació de vistes per la construcció dels mapes panoràmics.

El sistema proposat per Hirose dóna bons resultats en entorns estàtics; l'aparició d'elements mòbils en l'escena provocaria que aquests anessin apareixent i desapareixent de forma que l'usuari no tindria una percepció realística de l'entorn. La solució a aquest problema, que deixa proposada, inclouria un preprocessat més ampli de les imatges amb la detecció i tractament d'objectes mòbils, tal com es planteja recentment en treballs on es separen els objectes mòbils dels fons pre-modelats [Müller 05].

Figura 2.19 Sistema de representació i control emprat pel sistema de projecció de mapes panoràmics.

2.2.3.5 *Plenoptic Modelling*

A la Universitat de Carolina del Nord (UNC), Leonard McMillan i Gary Bishop treballen en la síntesi de noves imatges a partir de la reconstrucció de la funció plenòptica [McMillan-Bishop 95]. Aquesta funció, que va ser definida el 1991 per Adelson i Bergen del Massachusetts Institute of Technology, com la que recull tots els rajos de llum que arriben a un punt de vista des de tots els angles possibles. [Adelson-Bergen 91]. Primerament els autors demostren matemàticament que les condicions dels plans epipolars, es mantenen encara que les dades hagin estat capturades amb un model cilíndric. Aquesta serà la base per a la seva interpolació. Per a la captura de dades, prenen diferents vistes d'un escenari (properes entre elles) amb un sistema d'una camera que pot rotar 360º sobre un trípode [McMillan 95], només en el pla horitzontal. D'aquesta manera obtenen les vistes cilíndriques que usaran en l'emmagatzemament i càlculs posteriors (veure figura 2.20) [McMillan 99].

Figura 2.20 Captures de mapes panoràmics pel càlcul de la funció plenòptica.

Quan es demana la generació d'una imatge des d'un punt de vista diferent als enregistrats, es busquen les imatges més properes i amb rotacions i translacions es treuen les imatges candidates. Aquestes imatges són interpolades mitjançant algorismes d'alt nivell, que identifiquen regions semblants a les dues imatges (fent coincidir els contorns) i posteriorment interpolen els punts corresponents a les regions ocultes o de les quals no se'n té informació.

El mètode dóna uns resultats notables (veure figura 2.21) en escenaris sense moviment i on les dades s'han pres des de punts relativament propers (en l'experiment del que s'han extret les fotos la distància entre punts de vista era aproximadament de mig metre). Tant la síntesi de les noves imatges com la reprojecció de geometria cilíndrica a plana han estat realitzats amb una estació de treball SGI Indy sense acceleració de addicional en el maquinari, a temps propers al *video-rate* (5 imatges/segon).

Figura 2.21 Imatges sintetitzades per reprojecció de la geometria cilíndrica.

En els darrers anys, els esforços s'han dirigit principalment a accelerar la representació de les escenes en resolucions superiors a les tractades per McMillan i Bishop, que és de 320x240 píxels i a comprimir la ingent quantitat de dades a emmagatzemar seguint el mètode [Tong-Gray 97]. La tècnica de light field compression permet reduir molt significativament el volum de dades, establint un compromís entre la quantitat de dades i la precisió geomètrica considerada. Un estudi teòric de la Universitat de Stanford [Ramanathan-Girod 02] estableix les relacions exactes entre la capacitat de compressió i la pèrdua de precisió geomètrica.

2.2.3.6 Vídeos en múltiples perspectives

Seguint amb els autors que generen imatge sintètiques a partir de processat de les imatges originals, a la universitat de Taiwan, el grup de H.C Huang, C.C. Kao, Y.P. Hung i S.H. Nain genera vídeos en múltiple perspectiva, a partir de dues seqüències

d'entrada, això sí; per a cada parell de imatges preses imatge a imatge, generen els corresponents mapes de disparitat. Això té un cost computacional altíssim que els autors justifiquen [Huang 98] pel fet de que tot aquest processat pot realitzar-se *off-line* i no afecta el temps de còmput en el moment de la reproducció. Els mapes de disparitat són usats per interpolar les noves vistes en temps d'execució.

Figura 2.22 Generació del mapa de disparitat entre dues seqüències d'imatges per interpolar vistes en temps d'execució.

L'extensió del processament de vídeos en múltiples perspectives a càmeres en moviment ha estat utilitzat per Zhu, Riseman i Hanson [Zhu 01] per la creació de mosaics estereoscòpics on a més del problema geomètric cal enfrontar-se al "cosit" i alineament entre les diferents vistes d'una escena. Aquest mosaic (figura 2.23), no es crea com en altres treballs a partir dels mínims errors de veïnatge entre imatges sinó que s'empra la informació tridimensional inherent al mapa de disparitat per realitzar l'encaix entre les vistes, aplicacions com Google Earth i Microsoft Virtual Earth han popularitzat la utilització de mosaics generats amb fotos de satèl·lits, fotos aèries i informació tridimensional, obtinguda de bases de dades topogràfiques.

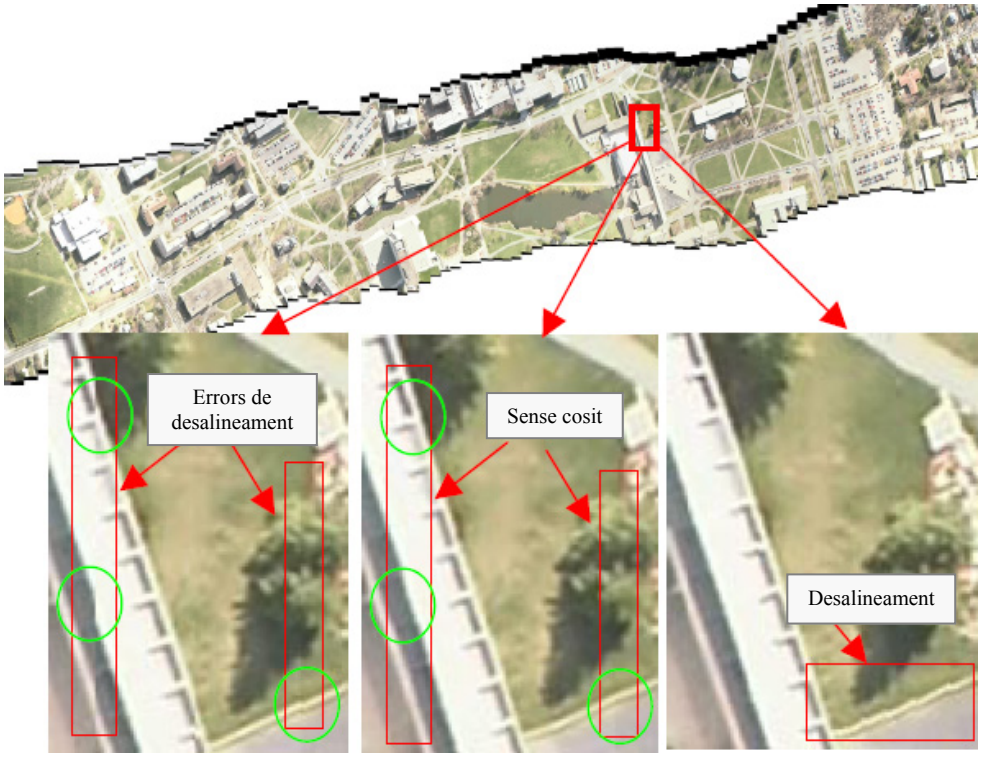


Figura 2.23 Generació de mosaics amb cosit entre vistes emprant la informació tridimensional del mapa de disparitat. Es mostra a l'esquerra i a la dreta el problema del desalineament, corregit a la vista central amb la informació de disparitat.

2.2.3.7 Representació de múltiples vistes amb disparitat usant compressors de vídeo mpeg-4

En aquest cas, els autors de l'institut Heinrich Hertz de Berlín, [Ohm-Müller 99] també fan captura de seqüències de vídeo des de múltiples perspectives, amb un entorn controlat on els objectes són fàcilment segmentables del fons i, a partir de les diverses vistes, extreuen els mapes de disparitat i les textures dels objectes de l'escena. Aquesta informació és llavors codificada en forma de seqüència d'imatges, amb un compressor de vídeo mpeg-4, on s'inclouen els tres canals convencionals R, G i B, i en el canal *alpha* reservat per imatges que incorporen graus de transparència, hi guarden la informació de disparitat. D'aquesta manera s'aprofita la capacitat dels compressors comercials de vídeo per guardar estructures de dades que després seran fàcilment utilitzades per generar les noves vistes dels objectes (figura 2.24). Això ho farà una aplicació interactiva descomprimint la informació de textura i la informació de disparitat i emprant-la per la interpolació de les noves vistes.

Figura 2.24 Utilització dels compressors comercials de vídeo per codificar els mapes de disparitat en el canal *alpha* de la seqüència. Les dues vistes superiors són les imatges capturades, la vista inferior dreta és el mapa de disparitat codificat al canal *alpha*, i la inferior esquerra la informació de textura codificada als canals R,G i B del vídeo.

Aquest mètode s'ha emprat en experiments de creació d'un canal televisiu amb selecció del punt de vista per part de l'espectador [Müller 02], i actualment, s'ha actualitzat el codificador de vídeo pel mpeg-7 [Müller 03] amb un increment en la qualitat de les imatges obtingudes amb el mateix volum d'informació.

2.2.3.8 Interpolació de vistes a partir d'una matriu de càmeres.

Es mostrarà ara un treball d'interpolació de vistes a partir d'una matriu de càmeres, que a diferència del presentat als apartats 2.2.2.1 i 2.2.2.2 no arriba a explicitar

la informació tridimensional obtinguda de l'escena. També difereix en la manera en que es situen les càmeres: mentre el treball de CMU construïa un *dome* (estructura semiesfèrica) de càmeres, aquí es disposaran en forma de matriu damunt d'un pla, cosa que facilitarà la geometria de la interpolació entre vistes.

Aquests treballs d'interpolació de vistes van ser iniciats pels enginyers de Canon l'any 1995 [Canon 95] i a la Universitat de Tokyo li han donat una robustesa teòrica i han permès determinar les limitacions del mètode i han fet les primeres proves amb imatges interpolades a temps real [Naemura-Harashima 99]. Inicialment aquests algoritmes d'interpolació són els més senzills i fàcils d'implementar per hardware degut a la menor complexitat dels processos que s'hi realitzen, que s'han d'executar a gran velocitat, i que són molt repetitius.

El treball dels investigadors de Canon es basa en les propietats de la geometria epipolar i la trajectòria prevista dels raigs de llum, suposant que la llum captada en una línia horitzontal concreta: x, de dues CCD de dues càmeres properes i alineades, ha de seguir un patró similar. La figura 2.25 mostra com la disposició de les càmeres permet la interpolació de noves vistes, associades a "càmeres virtuals". La unió en el pla definit com x-X (x de la línia de la imatge i X com a eix de coordenades absolut en el posicionament de les càmeres) dels segments corresponents als diferents objectes permet mitjançant un tall vertical, trobar la imatge sintetitzada en la línia concreta per una nova posició. Repetint aquest procés per totes les línies s'interpolarà la imatge sencera [Canon 95].

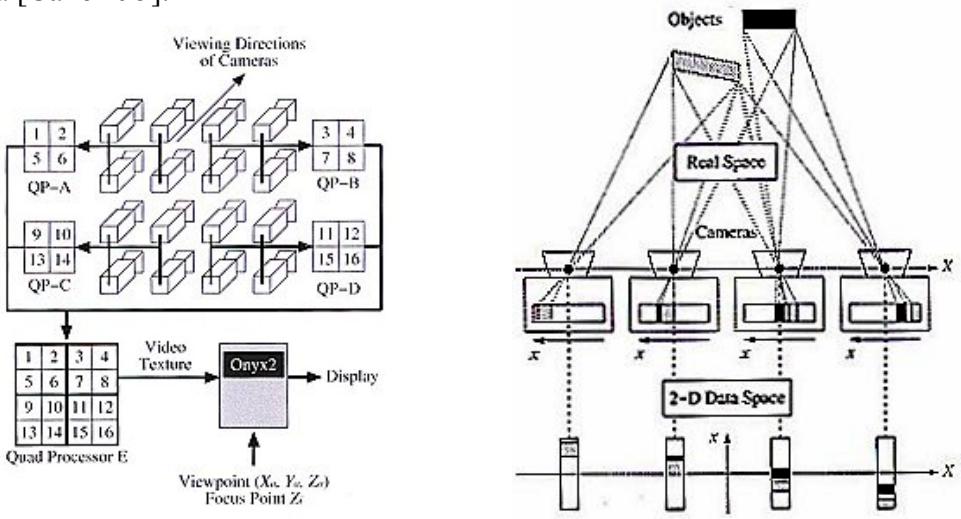


Figura 2.25 Distribució de les càmeres a l'espai i connexió als elements *quad-processor* (esquerra) i creació de l'espai bidimensional x-X que, representa una línia de la imatges damunt les diferents CCD.

Els enginyers de Canon van plantejar els quatre primers problemes d'aquest mètode que són: el *covering* o falta d'informació sobre objectes de la imatge que han quedat amagats per altres objectes, la *specularity* o existència de superfícies especulars en la imatge, la *degeneracy* o variacions en la informació del color durant la interpolació i el *deficit*, que implica la falta absoluta de informació en alguns punts concrets de la imatge [Canon 95]. Els problemes de *covering* i *specularity*, de la mateixa manera que en la majoria de mètodes mostrats anteriorment [Seitz 97] [McMillan-Bishop 95], no poden ser tractats amb la informació disponible i simplement s'eviten. La *degeneracy* i el *deficit* poden tractar-se per algoritme millorant paràmetres de la interpolació, fent reconstruccions locals de contorns, etc.

Per la seva banda, a la Universitat de Tokyo han donat un formalisme a aquest mètode d'interpolació, mitjançant la representació de les diferents imatges capturades en espais 5-Dimensionals (dimensions x i y dels píxels, X i Y en la distribució de les càmeres i L lluminositat captada per cada píxel) en els quals es fan interpolació sobre diferents subespais; així realitzen la interpolació tant en sentit horitzontal com vertical [Naemura-Harashima 99]. A l'hora de relacionar elements de cada parell de línies horitzontals, Naemura i Harashima proposen quatre possibles mètodes: el de *Nearest Neighbour*, que busca el segment més proper a la interpolació; el d'aproximació per plans, que dóna uns bons resultats excepte en les zones de superposició (veure discontinuïtats en la figura 2.26); el d'aproximació per fractals, que dóna bons resultats però té un altíssim cost de càlcul; i un mètode basat en l'estructura local, que dóna els millors resultats, ja que cerca característiques properes als objectes de la zona determinada a interpolar i tanca contorns, rehistograma els colors, etc.

La figura 2.26 mostra els resultats de la captura de les imatges d'una escena des de setze càmeres, i a temps real, generen una nova vista sintetitzada des d'una càmera virtual. Les setze imatges són capturades i incorporades a quatre *Quad-Processors*, que ajunten les imatges i passen el conjunt a un cinquè *Quad-Processor* que fusiona les setze imatges en una sola, que és passada a una *workstation* Onyx2. A partir d'aquí, per software es procedeix a la síntesi de noves vistes. El primer pas a realitzar per l'algoritme és alinear les setze imatges respecte a una referència comuna; per això s'extreu alguna característica comuna a totes les imatges (contorns, màxims....) i traslladant les imatges, es força la alineació entre elles. Seguidament s'interpola la imatge des de la nova càmera virtual amb l'algoritme descrit en sentit horitzontal i vertical; els resultats són mostrats a la figura 2.26.

El procés d'interpolació de vistes pot funcionar a temps real [Naemura-Harashima 99] treballant amb les setze imatges preses amb nivell de grisos i amb una resolució a la sortida 320x240 píxels. En l'exemple mostrat, l'algoritme d'interpolació de noves vistes no suavitza les discontinuïtats en la imatge resultat per un compromís amb la velocitat de procés.

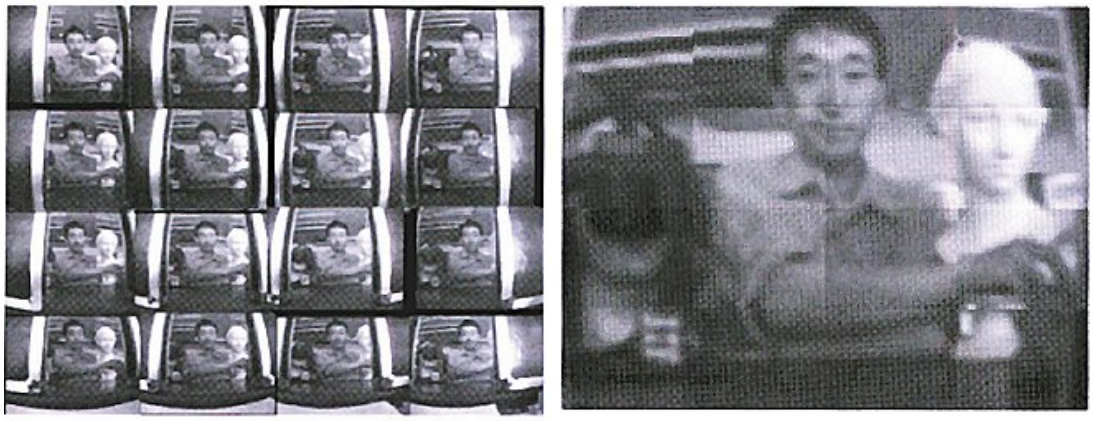


Figura 2.26 Captura de les setze imatges d'entrada al sistema (esquerra) i resultat de la imatge interpolada des d'un nou punt de vista, a temps real, i amb una resolució de 320x240 píxels.

L'experiment de Naemura i Harashima està concebut per permetre col·locar una càmera virtual en qualsevol posició del pla definit per les càmeres. Si es vol extendre la posició de la càmera a qualsevol ubicació fora del pla, o generalitzar la orientació de les càmeres, caldrà aplicar un mètode amb menys restriccions geomètrique, com és el de rectificació de tres vistes.

2.2.3.9 Síntesi per interpolació amb el mètode de rectificació de tres vistes.

El mètode de rectificació de tres vistes consisteix en, per dues vistes obtingudes a partir de càmeres reals i una de virtual, corresponent a on es vol obtenir la vista interpolada, fer la projecció de les tres vistes a un pla comú (veure figura 2.27) en que es podrà calcular la disparitat entre les vistes originals i fer la interpolació dels nous punts com en els mètodes anteriors [Scharstein 99].

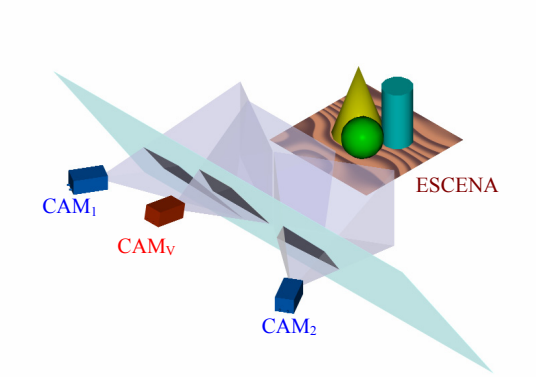


Figura 2.27 representació de la rectificació de tres vistes, amb la reprojecció de les imatges damunt d'un pla comú per afavorir la generació del mapa de disparitat i la interpolació de la nova vista.

Aquest procés de projecció de la imatge a un nou pla, es diu rectificació, i provocarà una deformació a la imatge homologable a la produïda per una operació de *warping* [Wolberg 90] [Medioni 04]. Aquesta transformació es pot veure a la imatge de la figura 2.28 [Criminisi 05]. En funció de la deformació realitzada caldrà interpolar alguns punts en la imatge resultant, com en el cas anterior i també, com en el cas anterior, serà necessari obtenir un bon mapa de disparitat per alguna de les tècniques de visió per ordinador. Deixant de banda la dificultat d'obtenir un mapa de disparitat suficientment dens, el mètode en conjunt segueix sent força costós i es segueix treballant en vies per obtenir les noves vistes a temps real i amb imatges de qualitat suficient [Lei-Hendriks 02].

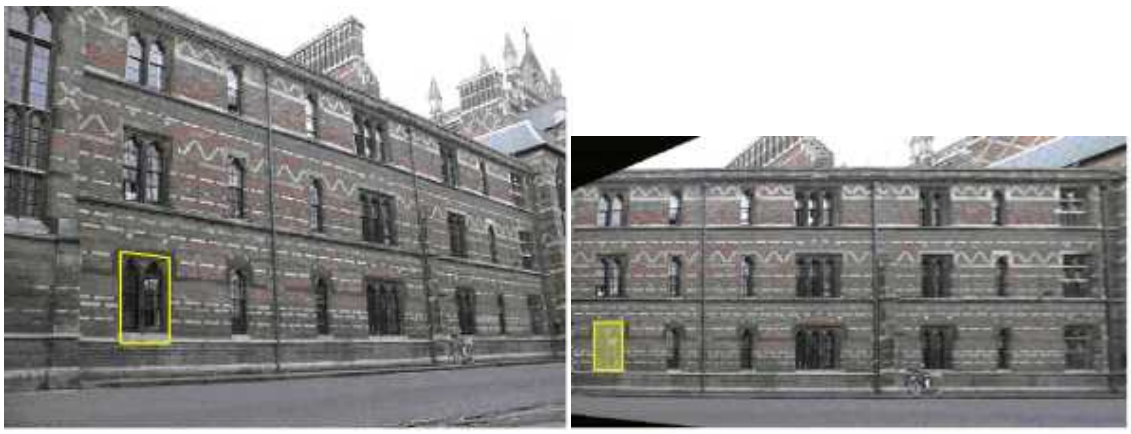


Figura 2.28 Mostra de la rectificació d'una vista a partir de la imatge original (esquerra) i una aplicació de quatre punts de la primera a la segona vista (rectangle en groc).

2.2.4 Taula resum dels mètodes de síntesi de vistes

La taula mostrada a continuació sols intenta donar, amb un cop d'ull, una idea del funcionament d'alguns dels mètodes presentats. La qualificació dels mètodes s'ha fet a partir dels comentaris dels propis autors en els articles referenciats i les dades presentades per ells.

Classificació emprada		Nom emprat pel mètode	Centre de recerca	Correctesa imatge	Ubicació cameres	Temps còmput
Treball només en 2D		View Morphing	University of Wisconsin	Depèn de l'objecte	No aplicable	Molt baix
Treball amb ús del 3D	Amb ús de 3D explícit	Virtualized Reality	CMU	Molt alta	Semicúpula amb càmeres al voltant de l'objecte	Baix per l'alt nombre de recursos emprats
		Mar de cameres	CMU-UNC	Alta	Idem.	Baix (però no interactiu)
		Equacions parametritzades	INRIA	Alta	Qualsevol amb configuració estèreo	Molt alt
		Generalized Voxel Color.	HP - Univ. Geòrgia	Alta	Al voltant de l'escena	Molt alt
	Amb ús implícit del 3D	Projecció de plans	UNC	Mitjana	Enfront de l'escena	Molt baix (interactiu)
		Models Projectius	INRIA	Alta	Enfront de l'escena	Alt
		JVT	INRIA	Alta	Enfront de l'escena	Alt
		Mapes Panoràmics	U. Tokio	Mitjana	Convexa	Baix
		Plenoptic Modelling	UNC	Alta	Convexa	Baix
		Múltiple perspectiva	National Taiwan Univ.	Mitjana	Qualsevol amb configuració estèreo	Molt baix (interactiu)
		Codificació vistes i disparitat en mpeg	HHI Berlín	Alta*	Qualsevol amb configuració estèreo	Baix
		Interpolació de vistes	Canon Inc.	Mitjana	Configuració estèreo epipolar	Molt baix (interactiu)
		VS amb rectificació	University of Cornell	Alta*	Qualsevol amb configuració estèreo	Baix

Taula 2.1 Taula resum amb les característiques de diversos mètodes de síntesi de vistes en funció de la utilització de la informació tridimensional de l'escena. *depenent de la qualitat del mapa de disparitat.

2.3 Reconstrucció tridimensional d'objectes

Com s'ha vist en l'apartat de síntesi de vistes, existeix una íntima relació entre la possibilitat d'obtenir bones noves vistes d'una escena i la qualitat i quantitat d'informació tridimensional obtinguda. L'extracció de la informació tridimensional d'una escena a partir d'una càmera i marques, dues càmeres, una càmera en moviment, o un conjunt de càmeres ha estat des de sempre el *leitmotiv* de la disciplina de la visió per ordinador.

No és la intenció d'aquesta breu referència en l'estat de l'art de la tesi, parlar de totes les tècniques amb que la visió per ordinador extreu informació tridimensional d'una escena, per això es pot consultar un llibre clàssic de visió per ordinador [Jain 95], la plana a *internet* de la Carnegie Mellon University dedicada a la visió per ordinador [CompVision 06] o algun llibre indicatiu de les noves tendències [Forsyth 02][Medioni 04]. Els mètodes vistos en el darrer apartat d'obtenció de la informació tridimensional, comprenen bàsicament les tècniques d'aparellament estèreo i els mètodes volumètrics.

- Es pot trobar una taxonomia extensa i completa dels mètodes de correspondència estèreo per obtenir mapes de disparitat densos en el treball de Scharstein i Zelisky [Scharstein 01].
- Els mètodes volumètrics habitualment reconstrueixen l'escena en base a la discretització de l'espai anomenada vòxel. Aquests vòxels, referits a un espai Euclidià, es poden organitzar eficientment en arbres tipus *octree* [Chien 86] [Potmesil 87] pels que s'han desenvolupat mètodes ràpids de selecció [Szeliski 93].
- Alternativament, alguns autors treballen amb l'estructura tridimensional creant un espai de disparitat (x, y, d, k), que per una càmera donada representa la disparitat d del píxel (x, y) amb qualsevol altra càmera k [Szeliski 99].
- El mètode de representació en plans [Baker 99], substitueix l'espai de disparitat per un conjunt de plans dependents de l'escena amb que es van aproximant les projeccions dels diferents objectes. Aquests plans a diferents profunditats, codifiquen la posició dels objectes segmentats de l'escena.
- Finalment, quan les càmeres no estan completament calibrades, es pot emprar una representació en espai projectiu [Kimura 99] [Saito 99] que condueix a l'anomenat conjunt de vòxels en espai projectiu.
- Com en el cas dels mètodes per obtenir mapes de disparitat, un report de l'Institut de Tecnologia de Geòrgia [Slabaugh 01] els dissecciona i explica els mètodes existents per la reconstrucció volumètrica a partir de fotografies.

Els mètodes volumètrics tendeixen a obtenir l'anomenada carcassa visual o *visual hull* d'un objecte [Laurentini 94] que és una aproximació del volum de l'objecte que segons les vistes triades i la geometria de l'objecte, aproxima millor la forma real de l'objecte que la clàssica carcassa convexa o *convex hull*. Les dues opcions principals per

obtenir aquest volum són els mètodes anomenats *space carving* i el *voxel coloring*. Convencionalment, el primer es basa en la projecció de les imatges damunt el volum de vòxels per triar els que s'han de mantenir i els que no [Kutulakos 00] i el segon en la projecció dels vòxels a les vistes per veure si els colors obtinguts pel vòxel són o no coherents [Seitz 99].

3. Obtenció de vistes per a l'experimentació

En aquest capítol es mostrarà breument el sistema físic emprat per a l'adquisició d'imatges dels objectes reals, s'explicarà com es pot guardar aquesta informació, la capacitat de disc necessària en funció del mètode de compressió i qualitat desitjada i, finalment, de quina manera es pot accedir eficientment a aquesta informació.

3.1. Obtenció de la informació fotomètrica d'un objecte.

Per l'elaboració dels experiments plantejats en aquesta tesi, és imprescindible disposar, en primer lloc, de la informació fotomètrica dels objectes que es voldran introduir posteriorment en un entorn de realitat augmentada o telepresència. Aquesta informació fotomètrica serà, bàsicament, imatges, vistes de l'objecte, preses des de ubicacions conegudes. Per a obtenir-les s'ha usat, naturalment, una càmera, un sistema d'adquisició de vídeo i un sistema robotitzat per moure la càmera o l'objecte amb precisió suficient.

Com a càmera s'ha usat una càmera amb estàndard de vídeo PAL (concretament un dispositiu Sony[1] FCB-75A), amb un captador tipus CCD de mitja polzada i una òptica amb *zoom* i enfocament motoritzat de entre 25 i 70mm. Com a dispositiu d'adquisició de vídeo s'ha usat una targeta d'un canal Euresys[1] Piccolo que captura dades RGB amb vuit bits de resolució per cada color. La tarja es connecta a un ordinador personal tipus PC, amb arquitectura intel[1] i sistema operatiu Windows XP[1] de Microsoft[1].

El sistema robotitzat s'ha construït expressament per als experiments realitzats. Donat que és un posicionador angular, amb dos graus de llibertat: capcineig (en l'estàndard anglès de robòtica *pitch*) i guinyada (en anglès *yaw*) es va decidir anomenar-lo projector gnomònic[2]. L'element terminal del robot serà la càmera que captura (on es projecta) la informació fotomètrica de l'objecte. Cal dir que els dos moviments es transmeten, un a l'objecte (la guinyada) i l'altre a la càmera (el capcineig). D'aquesta manera el sistema pot donar, d'un objecte, 1200 vistes en rotació respecte l'eix vertical, que s'anomenaran vistes en longitud, i 300 vistes en rotació respecte a l'eix horitzontal, que s'anomenaran latituds. La figura 3.1 mostra una projecció isonomètrica del disseny constructiu del sistema de projecció emprat.

[1] Sony, Euresys, Intel, Windows XP i Microsoft són marques registrades dels respectius propietaris.
[2] Del grec gnomon, veure glossari.

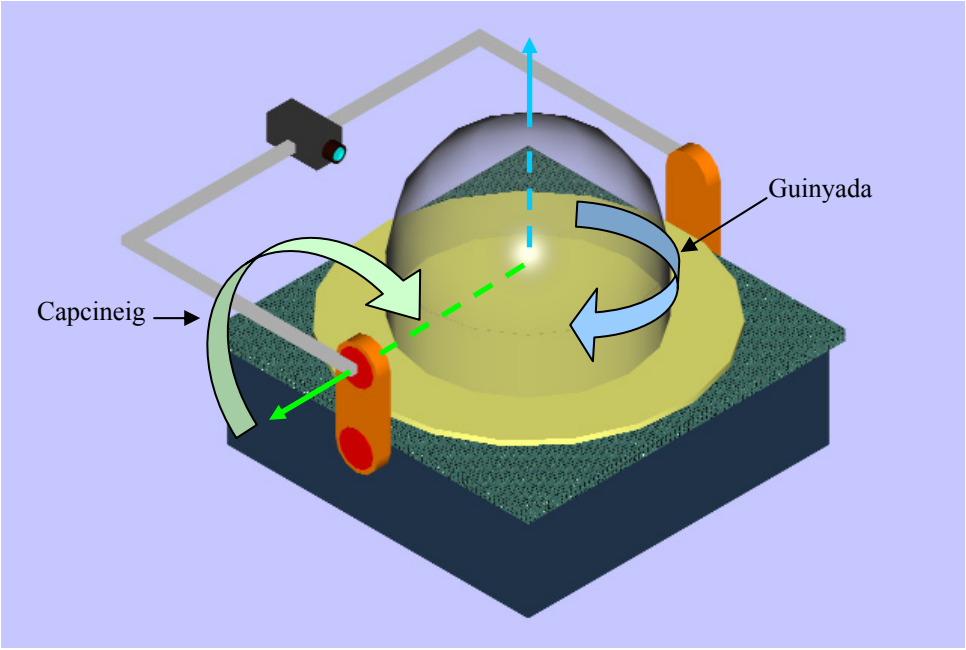


Figura 3.1 Definició dels moviments relatius a l'objecte: capcineig, respecte l'eix horitzontal (en verd), i ginyada, respecte al vertical (en blau).

El projector gnomònic emprat defineix l'espai útil disponible per a l'objecte com un cilindre de 18 cm de diàmetre, i 7 cm d'alçada coronat per una esfera de 9 cm de radi (veure figura 3.2). De l'objecte no s'obtindran, a priori, vistes des de sota, si es volgués fer amb aquest aparell, caldria fer-ho en dues passades, la segona amb l'objecte capgirat.

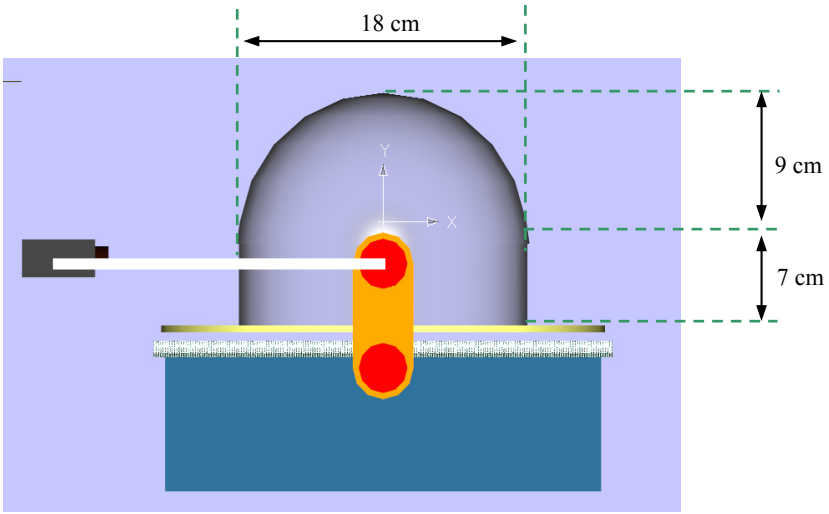


Figura 3.2 Espai de treball definit pel sistema d'adquisició robotitzat

3.2. Prestacions del sistema robotitzat. Estudi d'errors.

El posicionador emprat té una resolució horitzontal teòrica de 1200 passos per volta, i una resolució vertical teòrica de 300 passos. Donat que el sistema robotitzat s'ha equipat amb motors pas a pas i no s'ha tancat un llaç de control, S'ha cregut imprescindible fer un estudi de les resolucions reals obtingudes i dels errors comesos en

el desplaçament. Com a sistema de mesura s'ha emprat el mateix sistema d'adquisició, és a dir la càmera, per a determinar els errors del posicionador.

En primer lloc s'ha estudiat la resolució horitzontal del sistema. Per facilitar la detecció de les parts mòbils del robot mitjançant la càmera s'ha col·locat un patró mig blanc i mig negre (veure figura 3.3) enganxat al plat posicionador, de manera que es pugui fàcilment localitzar i determinar el píxel on es troba. La càmera ja s'havia calibrat prèviament amb un algoritme de visió per ordinador comú emprant llibreries específiques de l'eina matemàtica *matlab*[3].

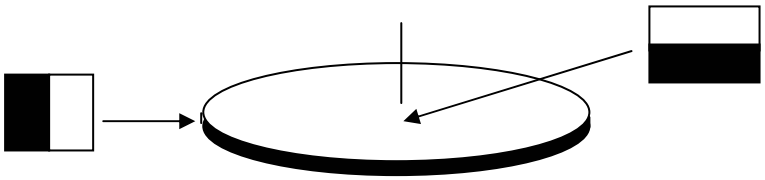


Figura 3.3 Patrons emprats per la calibració horitzontal i vertical del sistema.

Fent fer diverses voltes al sistema, detectant el pas cada vegada del patró horitzontal s'ha pogut determinar que la resolució mesurada per volta és de 1200,16 passos, és a dir, a cada sis voltes es perd un pas degut a la fricció del plat giratori. Estimant l'error de repetitivitat del sistema es pot concloure que per cada volta es tenen 1200,16 ± 0,16 passos. Considerant que els objectes es situen damunt d'un disc, s'obté que l'error de resolució lineal és de ($2 \cdot \pi \cdot$ radi / 1200,16).

Donat que el radi màxim acceptat pels objectes és de 9 cm, es tindrà una resolució de moviment horitzontal pitjor de 0,47 mm. Per un objecte de 6 cm de radi com el de les figures 3.5.a i 3.5.b s'obté una resolució en el moviment horitzontal de 0,31 mm. Donat que totes les mesures es fan relatives a una posició inicial, l'error de precisió no és massa important, de totes formes, si es volgués utilitzar el posicionador en mode absolut, es tindria un error de precisió de 0,3 mm pel radi de 9 cm i de 0,2 mm pel de 6 cm. Anàlogament, s'ha estimat la resolució vertical. Tot i que la teòrica és de 300 passos per volta, s'ha calculat una resolució mesurada de 300,75 passos. Això és degut als problemes mecànics al aixecar el braç robòtic. La repetitivitat calculada fa que es tingui una resolució de 300,75 ± 0,75 passos. Això fa que, en mil·límetres, s'obtingui un error de resolució de 0,31mm amb una precisió de 0,38 mm pel radi màxim de 9 cm i una resolució de 0,23 mm per un objecte de 6 cm de radi.

	Desplaçament horitzontal		Desplaçament vertical	
Tipus d'error	Objecte de 9cm de radi	Objecte de 6cm de radi	Objecte de 9cm de radi	Objecte de 6cm de radi
Error resolució	0,47 mm	0,31 mm	0,31 mm	0,23 mm
Error precisió	0,3 mm	0,2 mm	0,38 mm	0,27 mm

Taula 3.1 Errors de resolució i precisió en el desplaçament horitzontal i vertical del sistema del sistema projector.

La càmera emprada dóna una resolució de 768x576 píxels, dels quals la tarja d'adquisició en captura 720x576. Donat que la càmera té el *zoom* i l'enfocament motoritzats s'ha intentat sempre de maximitzar la superfície de l'objecte enquadrada.

[3] *Matlab* és un programari propietat de *Mathworks inc.*

Per un objecte de 18 cm de diàmetre, es té que cada píxel representa, a la part central de l'espai de treball, una longitud de 0,25 mm. Així doncs, per un objecte de mida màxima, es tindrà una resolució horitzontal aproximada d'un quart de mil·límetre. La resolució vertical en píxels serà de 576, dels quals s'intentarà que la majoria cobreixin la superfície de l'objecte. Per un objecte d'alçada màxima, de 16 cm, es tindrà que cada píxel representarà 0,3 mm damunt l'objecte.

Per un objecte com la urna funerària representada en la figura 3.5, que fa 12 cm de diàmetre per 14 cm d'alçada, i és quasi cilíndric, s'obté que cada píxel en les imatges representa 0,5 mm^2 damunt l'objecte.

La figura 3.4 mostra una fotografia del sistema d'adquisició construït, amb el PC emprat per l'adquisició i una decoració per poder segmentar fàcilment els objectes, que ofereix les prestacions que s'acaben d'enumerar.

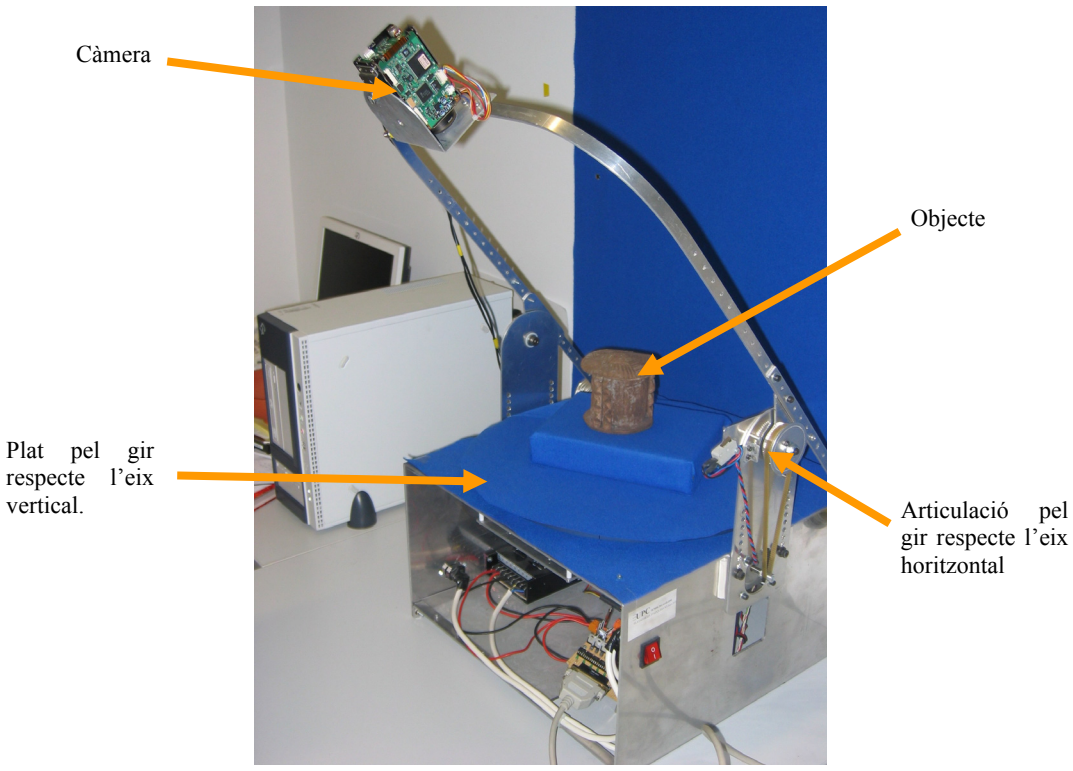


Figura 3.4 Robot posicionador emprat per a les captures.

Les imatges obtingudes per la càmera i la tarja d'adquisició tenen una resolució de 768 per 576 píxels, amb una profunditat de color de 24 bits, és a dir, 8 bits per cada canal de color primari: blau, verd i vermell. A la figura 3.5 es poden veure dos exemples del resultat de l'adquisició amb un mateix objecte en un recorregut vertical (una vista cada 10 graus de latitud) i un recorregut horitzontal (es mostra una vista per cada 10 graus de longitud). L'objecte mostrat és una urna funerària romana que s'ha situat damunt un fons de color blau per poder segmentar-la fàcilment amb algoritmes de selecció per croma (*chroma key*). L'objecte adquirit ha estat gentilment deixat per Sebastian Stride, del Departament d'Arqueologia i Història Antiga de la Universitat de Barcelona.

Figura 3.5.a Seqüència d'imatges en un recorregut respecte l'eix vertical (capcineig).

Figura 3.5.b Seqüència d'imatges en un recorregut respecte l'eix horitzontal (guinyada).

3.3. Emmagatzemament de vistes.

Donat que tant la càmera com el projector gnomètric estan connectats al PC que els controla i sincronitza, les imatges obtingudes es poden estructurar de forma natural com una seqüència on cada una de elles estarà presa des d'una posició coneguda del robot. Per cada objecte es tindrà doncs una taula de 1200*300 imatges, que representa totes les vistes disponibles del mateix.

El volum màxim de dades a manegar en principi serà de 1200x300x720x576x3 bytes, és a dir 447GB per objecte. Com que els sistemes operatius actuals ofereixen la capacitat d'emmagatzemar seqüències d'imatges (també conegudes com a *frames*) preses a intervals de temps constants com a fitxers de vídeo (extensions *.avi*, *.mpg*, *.qt*, *.wmv*), s'ha pensat en utilitzar aquests mateixos formats de fitxer per emmagatzemar seqüències d'imatges preses a intervals d'espai constants. Així, es podran aprofitar totes les funcionalitats dels fitxers de vídeo com ara emmagatzemament seqüencial indexat i compressió.

Els algoritmes de compressió es poden classificar entre els que tenen pèrdua de qualitat i els que no. Malgrat que el volum de dades resultant en els algoritmes de compressió sense pèrdua de qualitat no permet que els fitxers siguin portables ni fàcilment manejables per aplicacions posteriors de realitat augmentada, s'ha considerat imprescindible tenir-ne còpies en aquest format per poder avaluar la qualitat de les imatges generades pels diferents mètodes. La taula 3.2 mostra les mides a disc dels fitxers de seqüència de vídeo amb diferents formats de compressió, per uns quants objectes capturats amb el sistema.

Objecte Processat	Nombre de vistes	Mida sense comprimir	Compressió sense pèrdua	Compressor Cinepack	Compressor DivX	Compressor Xvid
	1200x300	447 GB	376 GB	130 GB	1,2 GB	0,9 GB
	1200x300	447 GB	375 GB	125 GB	1,1 GB	0,9 GB
	1200x300	447 GB	340 GB	120 GB	1 GB	0,9 GB
	1200x300	447 GB	380 GB	156 GB	1,4 GB	1,1 GB

Taula 3.2 Volum de dades resultant de l'aplicació de diferents mètodes de compressió en les seqüències d'imatges de diversos objectes: un *ciclamen* groc, urna funerària romana, una maqueta escala 1:20 d'un vehicle i una *viola cornuta*.

Els formats de compressió mostrats, normalment referenciats pel mot anglès *codec*, representen una selecció de les diferents opcions existents. Els mètodes de compressió sense pèrdua es basen en cercar les cadenes de bytes més repetides en els fitxers i substituir-les per les codificacions més curtes i les menys repetides per codificacions més llargues (mètodes amb codificacions de Hoffman) o en substituir els blocs repetits per la informació del bloc i el seu nombre de repeticions (codis RLE). *Cinepack* és un estàndard de compressió de vídeo clàssic molt utilitzat. Els formats DivX i Xvid s'han popularitzat els darrers anys amb les descàrregues de pel·lícules per internet, comprimeixen molt la informació i porten la pèrdua de qualitat fins al límit del que l'ull humà pot detectar. Avui en dia existeixen diversos dispositius comercials barats per accelerar la compressió i descompressió de vídeo en aquests formats. Els compressors amb pèrdua acostumen a basar-se en l'estàndard de compressió *mpeg*. Es pot trobar una especificació completa del funcionament dels algoritmes *mpeg* a [*mpeg* 06] però, de manera resumida es pot dir que el seu funcionament consisteix en:

1) Es divideix la imatge en N quadrats d'igual mida.
2) Si els píxels d'un quadrat donat no han sofert una variació suficient respecte a la imatge anterior, es manté l'anterior.
3) En els quadrats on s'ha detectat variació (moviment), s'aplica la transformada discreta del cosinus per obtenir una llista de coeficients.
4) Es guarden només els *x* primers coeficients, descartant els de més alta freqüència, no fàcilment apreciables per l'ull humà, però implicant certa pèrdua de qualitat.
5) Es crea una llista de coeficients de tots els quadrats on hi ha hagut moviment, que es comprimeix amb un algoritme d'informació sense pèrdua.
6) Es grava aquesta llista comprimida al disc, serà tota la informació necessària per reconstruir el *frame*.
7) Cada cert nombre d'imatges, es grava un frame complert (comprimit amb la transformada discreta del cosinus) per evitar que el procés de detecció de moviment indueixi massa errors.

Els valors dels paràmetres N, x i els llindar de variació d'un segment d'imatge respecte l'anterior determinen la qualitat i el percentatge de compressió obtingut.

3.4. Mètode d'accés a les vistes.

Un cop vist com es poden capturar i guardar les imatges obtingudes, sols queda mostrar el procediment que es pot emprar per a recuperar-les. Com s'ha vist les dades queden emmagatzemades en un fitxer de vídeo a disc, aquest fitxer és una seqüència indexada. Per la construcció del fitxer, els índexs representen la posició de la qual s'ha pres una vista determinada. Així doncs, si s'han pres M vistes per volta en un total de N voltes a diferents alçades consecutives, l'index $ind = (M \cdot j + i)$ representa la longitud i-èssima de la j-èssima latitud capturada.

Quan un sistema de realitat augmentada o de pintat d'objectes a distància demani la vista des d'un punt P, que està observant l'objecte situat en un punt C de l'espai, caldrà seguir la seqüència d'operacions següent:

- Obtenir la distància *d* a la que s'està observant l'objecte.
- Calcular els angles corresponents a la longitud i latitud observades.

- Indexar el fitxer amb els dos valors obtinguts, usant una funció d'accés directe al frame (les llibreries estàndard disposen de la funció *seek(frame)*).
- Aplicar un factor d'escala a la imatge en funció de *d* i de la distància càmera objecte del moment de la captura.
- Dibuixar la imatge al dispositiu corresponent.

Aquest mètode mostrat serà doncs una primera manera, senzilla, d'obtenir qualsevol vista d'un objecte des d'una aplicació interactiva. Consistirà en disposar d'un fitxer de vídeo on s'hagin gravat prèviament totes les vistes de l'objecte, organitzar-lo de manera que l'índex de les imatges del fitxer codifiqui la seva posició i, sota demanda, anar lliurant aquestes vistes a l'aplicació.

4. Síntesi de vistes

En aquest capítol s'explicarà en profunditat l'algoritme usat per, a partir d'un parell de vistes d'un objecte, interpolar-ne de noves, els problemes que planteja en certes situacions, i les millores que s'hi han incorporat.

4.1. Síntesi de noves vistes d'un objecte.

En primer lloc, es presentarà el mètode genèric d'interpolació de vistes, anomenat rectificació de tres vistes, per les dues reals i la virtual que es projecten en un pla per millorar les possibilitats d'interpolació.

4.1.1. Introducció.

El procés de síntesi de noves vistes a partir d'altres preexistents té com a objectiu generar imatges realístiques que no resultin artificioses a un observador humà. Per aconseguir-ho caldrà tenir el nou punt de vista en una zona relativament propera a les vistes de referència i si volem usar el mètode per a aplicacions com la tele-realitat caldrà garantir-ne l'eficiència.

El mètode seguit es basa en la *rectificació de tres vistes* i el *warping* (o deformació) d'imatges, tal com proposa Daniel Scharstein en el llibre "*View Synthesis Using Stereo Vision"* [Scharstein 99], sobre el que s'han aplicat certes millores tal com s'explicarà posteriorment. S'ha decidit triar aquest mètode perquè pot garantir la coherència física de les vistes generades i la seva formulació final permet una implementació ràpida de cara a aplicacions informàtiques interactives. En aquest mètode, les dues imatges originals i la imatge a sintetitzar són projectades i generada respectivament sobre un pla auxiliar per garantir una formulació senzilla del procés de síntesi (condicions epipolars). Un cop trobat aquest pla auxiliar només caldrà aplicar lleugeres deformacions a les imatges originals per a projectar-les en el pla, i al finalitzar, deformar la imatge sintetitzada perquè correspongui al punt de vista virtual.

Seguidament es parlarà de les condicions que han de complir els nous punts de vista, d'una millora proposada per a generalitzar la ubicació d'aquests punts, de la geometria implicada en la síntesi de les vistes, de l'ús dels mapes de disparitat i de la formulació final per a poder obtenir ràpidament i correctament les noves vistes a partir de les imatges originals.

4.1.2 El mètode de rectificació de tres vistes.

Els algoritmes basats en el mètode de *rectificació de tres vistes*, plantegen la situació mostrada en la figura 4.1, on tenim dues càmeres observant una escena i es volen obtenir la vista d'una càmera virtual situada entre les dues. Es pot establir aquí una primera condició que és que les càmeres reals i la virtual han d'estar mirant a la mateixa escena. Evidentment, si no existeix una zona estèreo o la càmera virtual no mira a la zona de la que tenim informació, no es podrà sintetitzar la nova vista. Una altra condició es que es suposa el sistema calibrat: coneixem els paràmetres extrínsecs (posició i orientació) i intrínsecs (bàsicament mides de la CCD, distància focal i alineament de la CCD amb la lent).

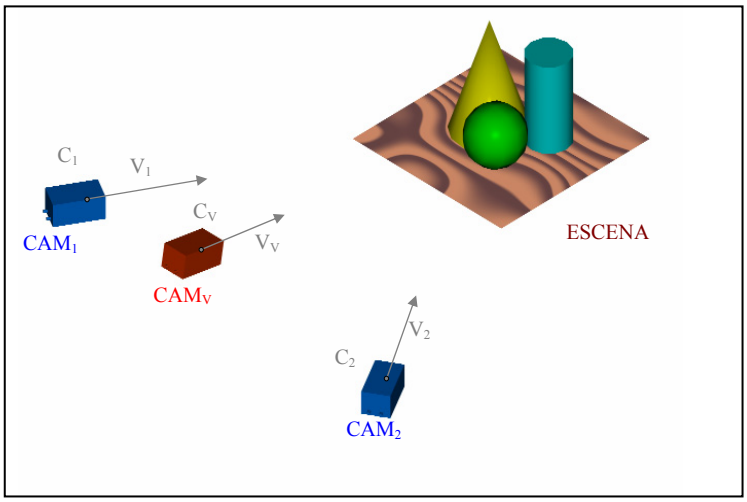


Figura 4.1 Disposició de les tres càmeres a l'espai.

Definirem els centres òptics de les càmeres reals com C_1 i C_2; i els vectors que defineixen els eixos òptics entre el centre de la CCD i el focus com V_1 i V_2 respectivament.

$$C_1 = \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} \quad , \quad C_2 = \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} \quad , \quad V_1 = \begin{bmatrix} v_{x1} \\ v_{y1} \\ v_{z1} \end{bmatrix} \quad , \quad V_2 = \begin{bmatrix} v_{x2} \\ v_{y2} \\ v_{z2} \end{bmatrix}. \qquad \text{(Def. 1)}$$

Per la càmera virtual definim el seu centre com C_V i V_V és el vecotr que defineix l'eix òptic.

$$C_V = \begin{bmatrix} x_V \\ y_V \\ z_V \end{bmatrix} \quad , \quad V_V = \begin{bmatrix} v_{xV} \\ v_{yV} \\ v_{zV} \end{bmatrix}. \qquad \text{(Def. 2)}$$

En aquest punt, el mètode planteja la reprojecció de les dues imatges reals i la virtual en construcció, sobre un pla paral·lel al format pels centres de les tres càmeres (figura 4.2). D'aquesta manera, s'obté una disposició en que les tres imatges són coplanars i per tant l'aparellament de punts de les imatges reals es podrà fer en

condicions de configuració epipolar i els píxels de la imatge virtual es podran obtenir per interpolació.

Dit d'altra manera, amb configuració epipolar, si un punt E de l'escena s'ha projectat a un píxel e_1 a la imatge de la càmera 1, es sap que la seva projecció e_2 a la imatge de la càmera 2 es troba sobre un segment rectilini determinat, i no cal buscar-lo en tota la imatge. A més, també es podrà veure que el punt E, vist des de la càmera virtual, es trobarà en un píxel e_v situat damunt la recta que uneix e_2 i e_1. Això implicarà una reducció de cost algorítmic en el moment de cercar els punts corresponents entre les dues vistes. D'altra banda. per fer la projecció d'una imatge IM_1 situada en el pla de la CCD a una nova imatge IM_1' situada en un nou pla, com pot ser P_R, sols cal conèixer la projecció de 3 punts [Wolberg 90] per construir una matriu de *warping* de la imatge.

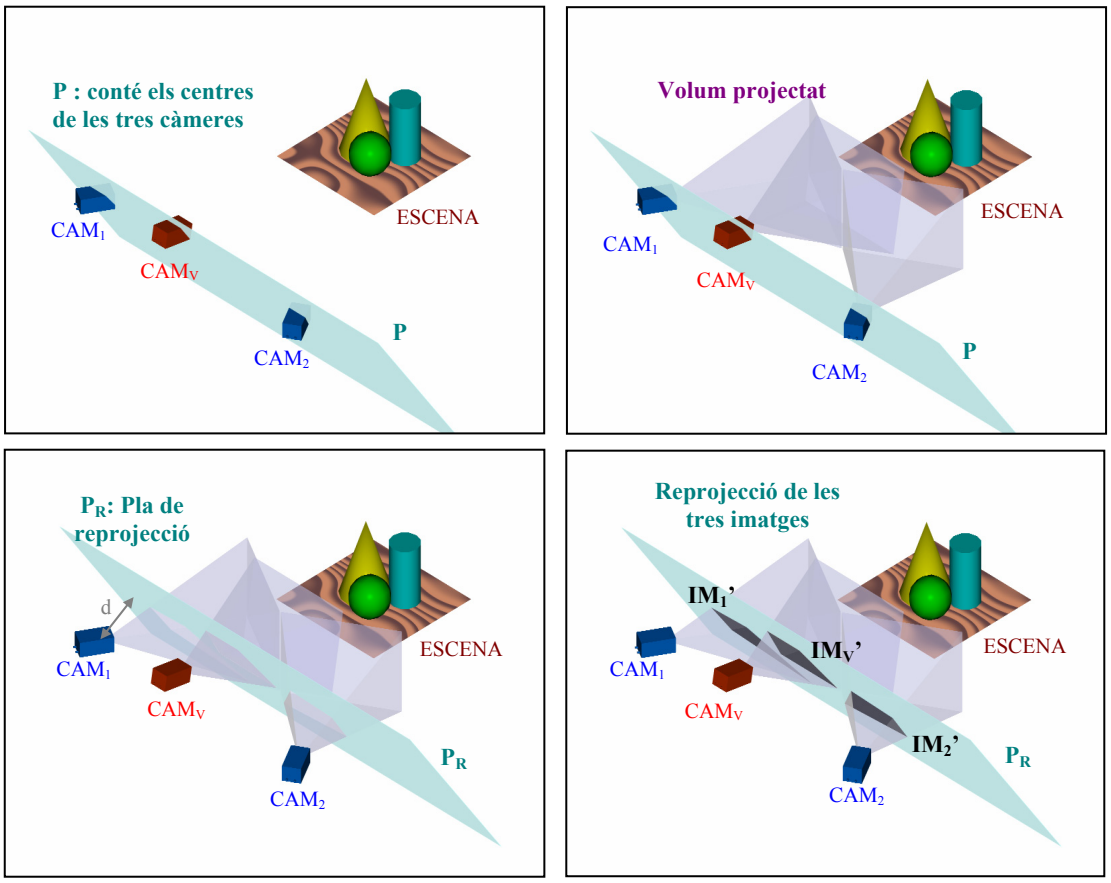


Figura 4.2 (d'esquerra a dreta i de dalt a baix): pla P que conté els centres de les tres càmeres; volums projectats per l'escena a cada una de les càmeres; pla P_R paral·lel al pla P situat a distància d; finalment, intersecció dels volums projectats amb P_R definint les tres noves imatges coplanars IM_1', IM_2' i IM_V'.

El càlcul del pla de reprojecció P_R es farà de la següent forma:

a) definim V_B com el vector de la *baseline* entre C_1 i C_2, i V_{AUX} és un vector auxiliar que uneix C_1 i C_V.

$$V_B = C_1 - C_2 = \begin{bmatrix} x_1 - x_2 \\ y_1 - y_2 \\ z_1 - z_2 \end{bmatrix} \quad , \quad V_{AUX} = C_1 - C_V = \begin{bmatrix} x_1 - x_V \\ y_1 - y_V \\ z_1 - z_V \end{bmatrix} . \qquad \text{(Def. 3)}$$

b) Cerquem V_P que serà el vector perpendicular al pla P que conté els centres de les tres càmeres.

$$V_P = V_B \times V_{AUX} = \begin{vmatrix} x_u & y_u & z_u \\ x_1 - x_2 & y_1 - y_2 & z_1 - z_2 \\ x_1 - x_V & y_1 - y_V & z_1 - z_V \end{vmatrix} = \begin{bmatrix} V_{PX} \\ V_{PY} \\ V_{PZ} \end{bmatrix}. \qquad \text{(Eq. 1)}$$

c) En aquest moment trobarem l'equació del pla P usant la definició del pla i solucionant K per qualsevol dels centres de les càmeres.

$$P: \quad X \cdot V_{PX} + Y \cdot V_{PY} + Z \cdot V_{PZ} + K = 0 . \qquad \text{(Eq. 2)}$$

d) Ara podem calcular l'equació del pla de reprojecció P_R. Donat que serà paral·lel a P, podem usar la mateixa equació (Eq. 2) variant el valor de K. Així doncs, solucionarem l'equació 2 per un punt situat a distància d del pla P. Aquest punt, C_P el calcularem així:

$$C_P = C_1 + d \cdot \frac{V_P}{\| V_P \|} . \qquad \text{(Eq. 3)}$$

De manera que, finalment, trobarem el pla de reprojecció P_R:

$$P_R: \quad X \cdot V_{PX} + Y \cdot V_{PY} + Z \cdot V_{PZ} + K_R = 0 . \qquad \text{(Eq. 4)}$$

Aquesta és la formulació estàndard del mètode de rectificació de tres vistes, per al càlcul del pla de reprojecció. A continuació es mostraran certs problemes detectats en el mètode, i les millores proposades per solucionar-los [Martín 03].

4.2. Millores del mètode

A continuació es presenta, pel mètode de rectificació de tres vistes, la identificació d'un parell de problemes en la seva formulació i les solucions proposades.

4.2.1 Construcció del pla de reprojecció.

El problema:

Algunes vegades, els centres de les tres càmeres queden en un pla paral·lel o quasi paral·lel a l'eix òptic d'alguna de les càmeres CAM_i, en aquest cas és impossible projectar la imatge IM_i sobre el pla degut a que el volum de projecció de la càmera té una intersecció infinita amb el pla.
Aquesta idea es mostra a la figura 4.3, on cada fila representa el procés de construcció del pla P a la primera vinyeta, mostrant el vector normal al pla i el vector de l'eix òptic

d'una càmera. A la segona vinyeta es mostra el pla de reprojecció P_R paral·lel a P. Finalment es mostra a la tercera vinyeta de cada fila es mostra el volum de projecció de la càmera i el pla de reprojecció. Les successives files mostren diferents ubicacions de la càmera virtual on cada cop és més difícil avaluar la intersecció entre el volum i el pla, fins a fer-se matemàticament impossible.

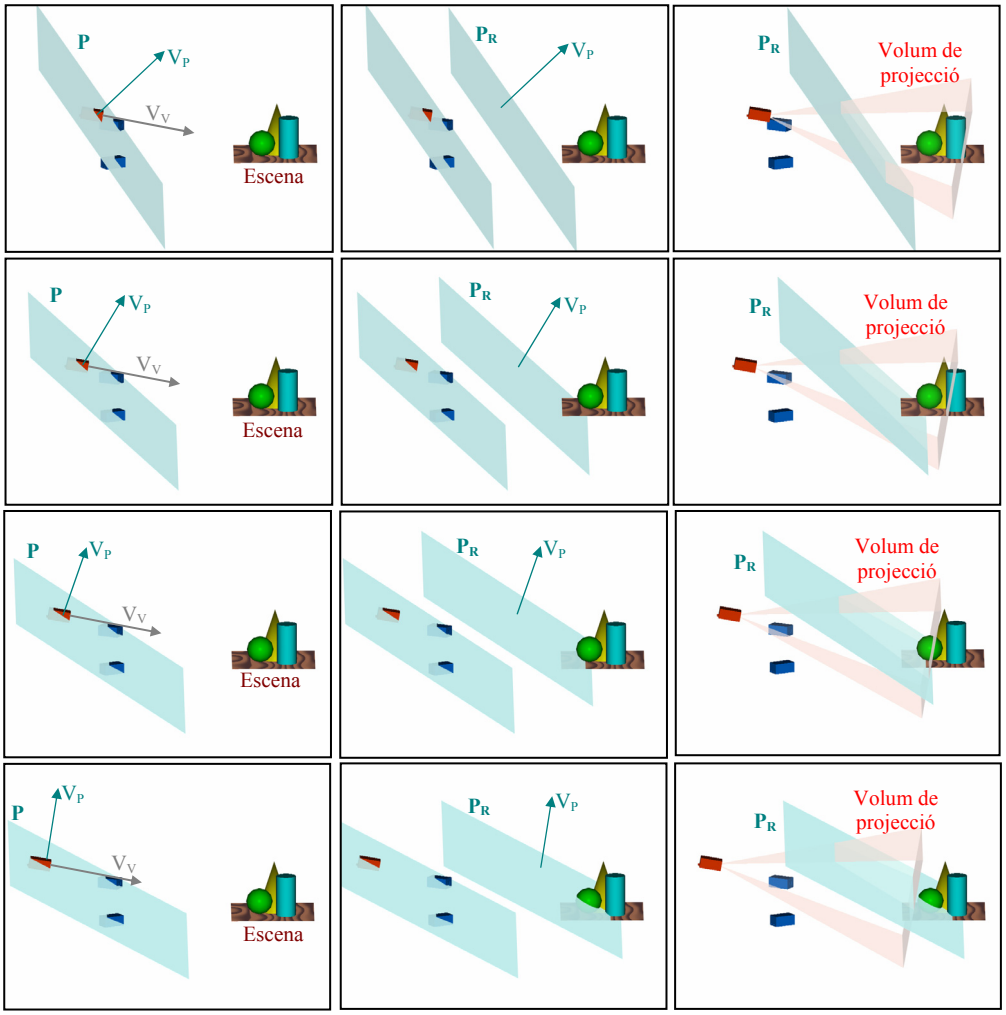


Figura 4.3. Representació en una sèrie de vinyetes de la impossibilitat de calcular la intersecció entre el pla P_R i el volum projectat a una càmera (per exemple, la càmera virtual) si l'eix òptic d'aquesta i el pla que conté els centres són quasi paral·lels.

Proposta de solució:

Per resoldre el problema presentat, donarem un grau de llibertat a la posició de la càmera de la que no es pot reprojectar la imatge; per exemple, seguint el mostrat a la figura 4.3, la càmera virtual. D'aquesta manera considerarem com a nou centre de la càmera virtual un punt anomenat C_S, en comptes de C_V. Aquest punt C_S es definirà de la següent manera:

$$C_S = C_V + \mu \cdot V_V \ . \qquad \text{(Eq. 5)}$$

Així, el centre de la càmera virtual es podrà moure en la direcció de l'eix òptic, endavant i endarrera, habilitant-nos per trobar un pla P, i el seu paral·lel per la reprojecció P_R, que pugui incloure les tres imatges. La figura 4.4 mostra com aquesta

idea ens permetrà solucionar el problema exposat. Amb aquesta modificació es podrà trobar la vista des de la nova ubicació C_S, per obtenir la vista desitjada des de C_V caldrà aplicar una matriu d'escalat entre les imatges.

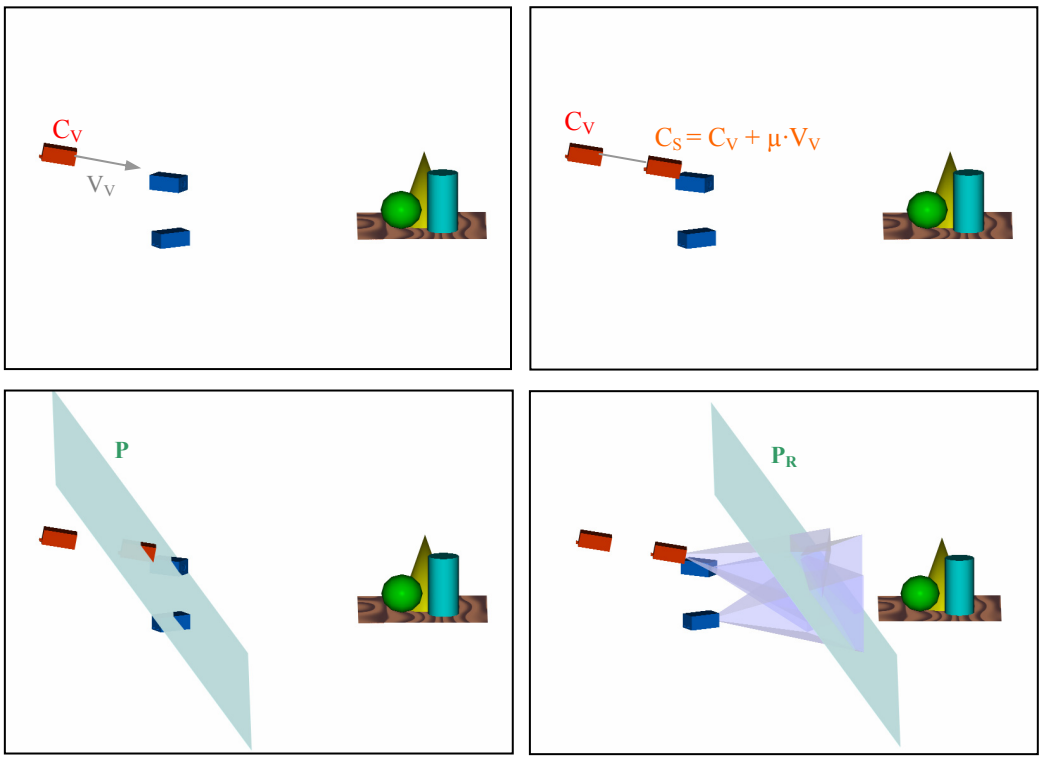


Figura 4.4 La primera vinyeta (d'esquerra a dreta, de dalt a baix) mostra la càmera virtual i el vector de l'eix òptic, la segona, mostra la nova ubicació possible per la càmera virtual C_S, la tercera mostra la construcció del pla P i la darrera el pla de reprojecció que ara si, pot contenir les interseccions amb els volums de projecció de les càmeres.

Ara, es pot cercar el valor òptim pel paràmetre μ que faci millor la projecció de les tres imatges sobre el pla P_R. Primer, es cercarà el pla P que conté C_1, C_2 i C_S, tal com s'ha descrit prèviament. Com sempre, es definirà el pla pel seu vector perpendicular V_P:

$$V_B = C_1 - C_2 = \begin{bmatrix} x_1 - x_2 \\ y_1 - y_2 \\ z_1 - z_2 \end{bmatrix} \quad , \quad V_{AUX} = C_1 - C_S = \begin{bmatrix} x_1 - (x_V + v_{xV} \cdot \mu) \\ y_1 - (y_V + v_{yV} \cdot \mu) \\ z_1 - (z_V + v_{zV} \cdot \mu) \end{bmatrix}. \qquad \text{(Eq. 6)}$$

$$V_P = V_B \times V_{AUX} = \begin{vmatrix} x_u & y_u & z_u \\ x_1 - x_2 & y_1 - y_2 & z_1 - z_2 \\ x_1 - (x_V + v_{xV} \cdot \mu) & y_1 - (y_V + v_{yV} \cdot \mu) & z_1 - (z_V + v_{zV} \cdot \mu) \end{vmatrix} = \begin{bmatrix} V_{PX} \\ V_{PY} \\ V_{PZ} \end{bmatrix}.$$

On:

$$V_{PX} = (z_1 y_v + z_2 y_1 + y_2 z_v - y_1 z_v - z_2 y_v - y_2 z_1) + \mu \cdot (y_2 v_{zV} + z_1 v_{yV} - y_1 v_{zV} - z_2 v_{yV}).$$
$$V_{PY} = (x_1 z_v + x_2 z_1 + z_2 x_v - z_1 x_v - x_2 z_v - z_2 x_1) + \mu \cdot (z_2 v_{xV} + x_1 v_{zV} - z_1 v_{xV} - x_2 v_{zV}).$$
$$V_{PZ} = (y_1 x_v + y_2 x_1 + x_2 y_v - x_1 y_v - y_2 x_v - x_2 y_1) + \mu \cdot (x_2 v_{yV} + y_1 v_{xV} - x_1 v_{yV} - y_2 v_{xV}).$$

(Eq. 7)

D'aquesta manera, s'ha obtingut el vector V_P parametritzat per μ; l'equació 7 ens ha definit $V_P(\mu)$ segons la qual podrem obtenir el millor valor de μ per evitar el problema de la projecció no calculable.
Per obtenir un bon pla de reprojecció es necessita que els vectors V_1, V_2 i V_V siguin molt similars al vector $V_P(\mu)$. Així s'utilitzarà l'angle entre els vectors com a mesura de la seva similitud. Recordant l'equació del producte escalar de dos vectors i definint els vectors unitaris de V_1, V_2, V_V i $V_P(\mu)$, respectivament com a V_{1u}, V_{2u}, V_{Vu} i $V_{Pu}(\mu)$, es poden fer les següents operacions:

$$V_{1u} \cdot V_{Pu}(\mu) = \cos(\alpha_1) \quad ; \quad V_{2u} \cdot V_{Pu}(\mu) = \cos(\alpha_2) \quad ; \quad V_{Vu} \cdot V_{Pu}(\mu) = \cos(\alpha_V) .$$

$$V_X \cdot V_Y = \|V_X\| \cdot \|V_Y\| \cdot \cos(\alpha) \quad \rightarrow \quad \frac{V_X}{\|V_X\|} \cdot \frac{V_Y}{\|V_Y\|} = \cos(\alpha) \qquad \text{(Eq. 8)}$$

On α_1, α_2 i α_V són, respectivament, els angles entre els vectors V_1, V_2, V_V i el vector perpendicular al pla de reprojecció V_P. Si aquests angles són propers a zero, el procés de rectificació de tres vistes podrà funcionar correctament, en canvi si els seus valors són propers a 90 graus, serà matemàticament impossible trobar les imatges reprojectades. Per tant, es té l'expressió dels tres angles parametritzats també segons μ i si es vol que els valors de α_1, α_2 i α_V siguin propers a 0, això vol dir doncs que el seu cosinus serà proper a 1; i 1 és el valor màxim de la funció cosinus; així si es defineixen les funcions F_1, F_2 i F_V, com:

$$\cos(\alpha_1) = \frac{V_1}{\|V_1\|} \cdot \frac{V_P(\mu)}{\|V_P(\mu)\|} \quad \rightarrow \quad F_1 := \frac{V_1}{\|V_1\|} \cdot \frac{V_P(\mu)}{\|V_P(\mu)\|} ,$$

$$\cos(\alpha_2) = \frac{V_2}{\|V_2\|} \cdot \frac{V_P(\mu)}{\|V_P(\mu)\|} \quad \rightarrow \quad F_2 := \frac{V_2}{\|V_2\|} \cdot \frac{V_P(\mu)}{\|V_P(\mu)\|} ,$$

$$\cos(\alpha_1) = \frac{V_V}{\|V_V\|} \cdot \frac{V_P(\mu)}{\|V_P(\mu)\|} \quad \rightarrow \quad F_V := \frac{V_V}{\|V_V\|} \cdot \frac{V_P(\mu)}{\|V_P(\mu)\|} .$$

(Eq. 9)

i s'intenten maximitzar independentment, s'obtindran tres solucions diferents per μ segons F_1, F_2 i F_V. Com que V_1, V_2 i V_V seran en general, vectors independents entre ells, aquestes solucions també seran incompatibles entre elles. Per tant, el millor serà buscar el valor per μ que maximitzi, en promig, els tres cosinus, i conseqüentment, apropi els tres angles cap a zero. Per aconseguir-ho es definirà la funció F com:

$$F := \frac{\cos(\alpha_1) + \cos(\alpha_2) + \cos(\alpha_3)}{3} . \qquad \text{(Eq. 10)}$$

de la que es cercarà el màxim segons el paràmetre μ. Això és:

$$MAX(F) = MAX\left(\frac{\cos(\alpha_1) + \cos(\alpha_2) + \cos(\alpha_3)}{3}\right) = MAX\left(\cos(\alpha_1) + \cos(\alpha_2) + \cos(\alpha_3)\right) =$$

$$MAX\left(\frac{V_1}{\|V_1\|} \cdot \frac{V_P(\mu)}{\|V_P(\mu)\|} + \frac{V_2}{\|V_2\|} \cdot \frac{V_P(\mu)}{\|V_P(\mu)\|} + \frac{V_V}{\|V_V\|} \cdot \frac{V_P(\mu)}{\|V_P(\mu)\|}\right) = MAX\left(\frac{V_P(\mu)}{\|V_P(\mu)\|} \cdot (V_{1u} + V_{2u} + V_{Vu})\right)$$

(Eq. 11)

De l'equació 7, es pot definir:

$$V_{PX} = A_0 + \mu \cdot A_1 \ , \quad V_{PY} = B_0 + \mu \cdot B_1 \ , \quad V_{PZ} = C_0 + \mu \cdot C_1 \ .$$

considerant com a constants els valors de A_0, A_1, B_0, B_1, C_0 i C_1 que seran :

$$\begin{aligned}
A_0 &= (z_1 y_v + z_2 y_1 + y_2 z_v - y_1 z_v - z_2 y_v - y_2 z_1), \\
A_1 &= (y_2 v_{zV} + z_1 v_{yV} - y_1 v_{zV} - z_2 v_{yV}), \\
B_0 &= (x_1 z_v + x_2 z_1 + z_2 x_v - z_1 x_v - x_2 z_v - z_2 x_1), \\
B_1 &= (z_2 v_{xV} + x_1 v_{zV} - z_1 v_{xV} - x_2 v_{zV}), \\
C_0 &= (y_1 x_v + y_2 x_1 + x_2 y_v - x_1 y_v - y_2 x_v - x_2 y_1), \\
C_1 &= (x_2 v_{yV} + y_1 v_{xV} - x_1 v_{yV} - y_2 v_{xV}) .
\end{aligned}$$

i definint els també constants A_V, B_V i C_V que seran:

$$\begin{aligned}
A_V &= \frac{x_1}{\sqrt{x_1^{\,2} + y_1^{\,2} + z_1^{\,2}}} + \frac{x_2}{\sqrt{x_2^{\,2} + y_2^{\,2} + z_2^{\,2}}} + \frac{x_V}{\sqrt{x_V^{\,2} + y_V^{\,2} + z_V^{\,2}}}, \\
B_V &= \frac{y_1}{\sqrt{x_1^{\,2} + y_1^{\,2} + z_1^{\,2}}} + \frac{y_2}{\sqrt{x_2^{\,2} + y_2^{\,2} + z_2^{\,2}}} + \frac{y_V}{\sqrt{x_V^{\,2} + y_V^{\,2} + z_V^{\,2}}}, \\
C_V &= \frac{z_1}{\sqrt{x_1^{\,2} + y_1^{\,2} + z_1^{\,2}}} + \frac{z_2}{\sqrt{x_2^{\,2} + y_2^{\,2} + z_2^{\,2}}} + \frac{z_V}{\sqrt{x_V^{\,2} + y_V^{\,2} + z_V^{\,2}}} .
\end{aligned}$$

es pot recuperar l'equació 11 rescrivint-la així:

$$MAX(F) \ = \ MAX\left(\frac{(A_0 + \mu \cdot A_1) \cdot A_V + (B_0 + \mu \cdot B_1) \cdot B_V + (C_0 + \mu \cdot C_1) \cdot C_V}{\sqrt{(A_0 + \mu \cdot A_1)^2 + (B_0 + \mu \cdot B_1)^2 + (C_0 + \mu \cdot C_1)^2}} \right)$$

Calculant la derivada de F segons la variable μ, es troba una única solució corresponent a un màxim de la funció en el punt μ_{MAX}:

$$\mu_{MAX} = \frac{NUM}{DEN} ; \qquad \text{(Solució 1)}$$

$$\begin{aligned}
NUM = {} & C_0^{\,2}(A_1A_2 + B_1B_2) + B_0^{\,2}(A_1A_2 + C_1C_2) + A_0^{\,2}(B_1B_2 + C_1C_2) - A_0C_0(C_1A_2 + A_1C_2) - \\
& B_0C_0(B_1C_2 + C_1B_2) - A_0B_0(B_1A_2 + A_1B_2) \\
DEN = {} & C_1^{\,2}(A_0A_2 + B_0B_2) + B_1^{\,2}(A_0A_2 + C_0C_2) + A_1^{\,2}(B_0B_2 + C_0C_2) - A_0A_1(B_1B_2 + C_1C_2) - \\
& B_0B_1(A_1A_2 + C_1C_2) - C_0C_1(A_1A_2 + B_1B_2)
\end{aligned}$$

Aquesta solució permetrà obtenir la millor posició per la càmera virtual. A partir d'aquest moment es podrà aplicar normalment el mètode de rectificació de tres vistes.

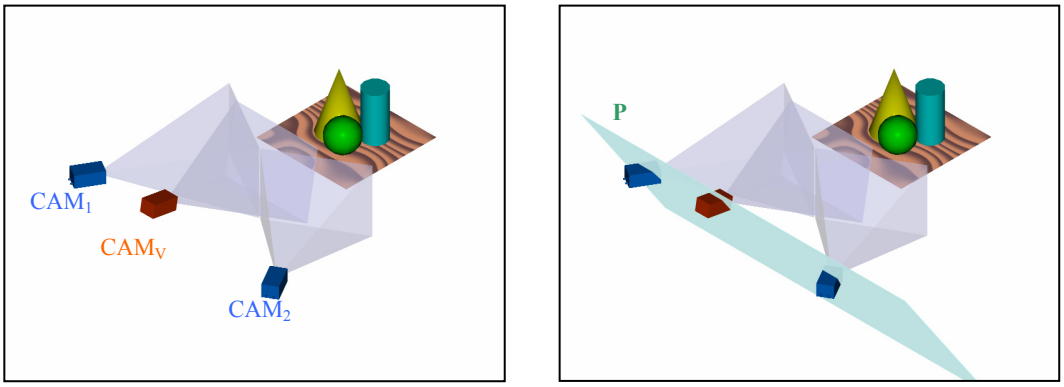
4.2.2 Distància del pla de reprojecció.

La distància entre el pla que conté els centres de les càmeres i el pla de reprojecció P_R (veure definició de *d* a la figura 4.2) es defineix habitualment com la mateixa distància que existeix entre els centres de les dues càmeres reals. Com es veurà, això dóna alguns avantatges matemàtics ja que els autors (Scharstein) defineixen un nou sistema de coordenades centrat a C_1, amb distància unitària definida com l'existent entre els dos centres de les càmeres reals. Això voldrà dir que en el nou espai, una càmera està a (0,0,0) l'altra a (1,0,0) i el pla de reprojecció és el definit Z=1, la qual cosa simplificarà el càlcul de les matrius de projecció i altres càlculs.

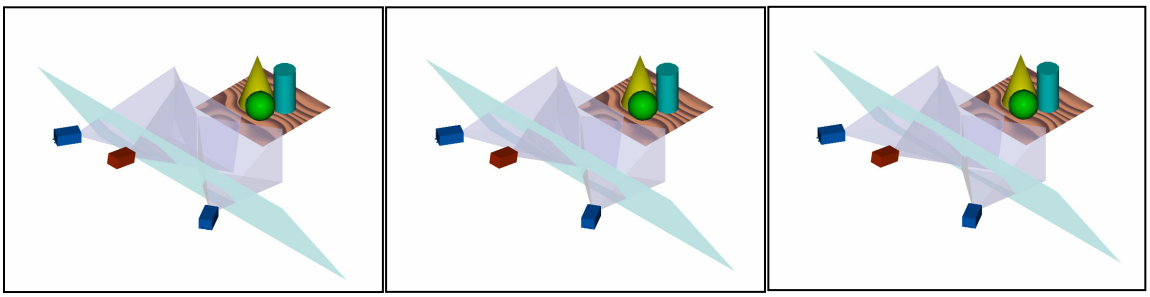
El problema:

S'ha observat que, algunes vegades, un valor per la distància entre els dos plans pot ser millor que un altre, en el sentit de que la imatge sintetitzada tindrà més qualitat o, simplement, més píxels que en un altre cas. Això es pot veure amb un exemple, que es detalla a continuació:

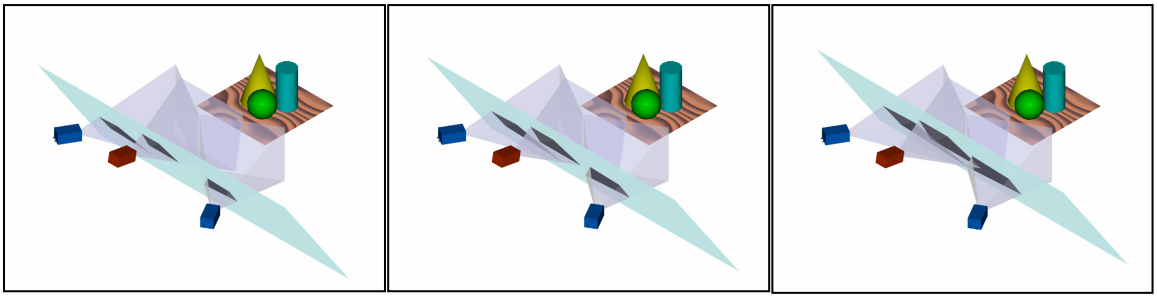
1) Primer es mostren les tres càmeres: CAM_1, CAM_2 i CAM_V i el pla que conté els centres de les tres càmeres: P.



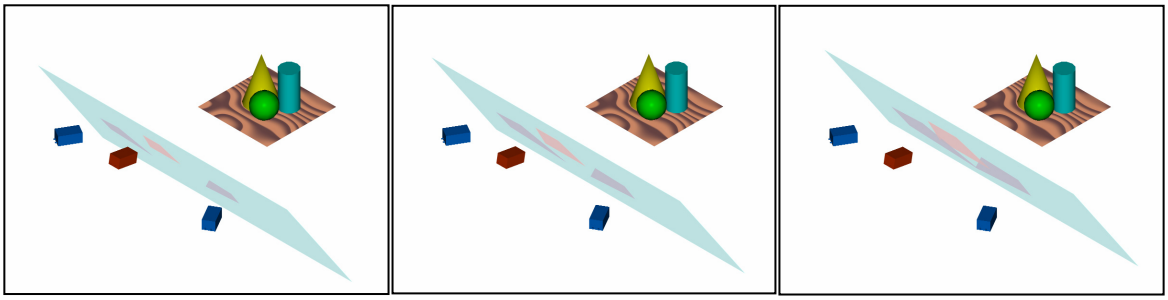
2) En segon lloc, es calcula el pla de reprojecció P_R, paral·lel a P, i situat a una distància *d* de P; es mostren tres possibles valors per d: d_1, d_2 i d_3 i els seus respectius plans de reprojecció:



3) Després es mostren les interseccions entre aquests plans i els volums de projecció de les càmeres per cada una de les distàncies considerades:



4) Finalment, es poden observar els resultats: la intersecció entre els volums de projecció i els plans de reprojecció a diferents distàncies generen les imatges IM_1', IM_2' i IM_V', amb aspecte divers segons la distància seleccionada.



L'algoritme de síntesi de vistes genera la imatge virtual IM_V' a partir de la informació donada per IM_1' i IM_2'. Es busca per a cada píxel de la imatge IM_1' el seu corresponent a IM_2' i es calcula la posició del punt real representat que es projecta a la imatge virtual. Amb la nova geometria i les imatges en el mateix pla P_R es sap que el punt corresponent a un donat es troba en una línia recta i aquesta és paral·lela a la que uneix les projeccions de C_1 i C_2. Per tant, es necessita que les tres imatges reprojectades estiguin el màxim d'alineades entre elles.

La següent figura mostra aquest alineament i com, la quantitat d'imatge virtual que es podrà generar dependrà de les distàncies seleccionades: d_1, d_2 i d_3.

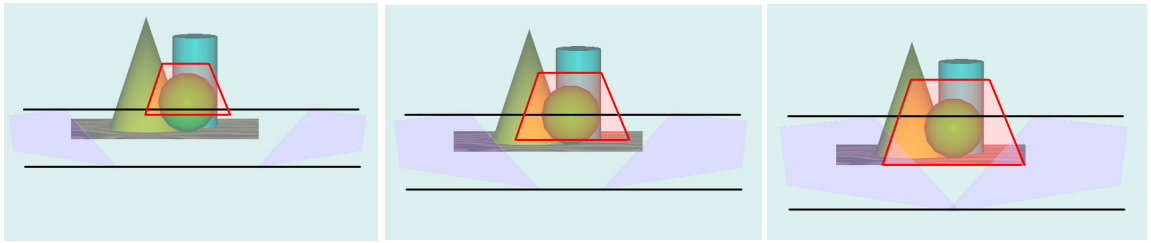


Figura 4.5 Projecció de les imatges reals (en blau) i virtual (en vermell) sobre el pla de reprojecció. També es veuen les línies d'alineament entre les imatges reals que defineixen l'àrea on es podran interpolar píxels a la imatge virtual. D'esquerra a dreta: per distàncies d_1, d_2 i d_3.

Com es pot veure, les dues línies horitzontals determinen l'àrea en la que es podrà interpolar la imatge virtual. Els trapezoides mostren les projeccions de les imatges reals i virtual sobre el pla. En el tercer cas, la imatge que es pot sintetitzar serà major

que en els altres, per tant, una bona selecció de la distància *d* de projecció permetrà tenir una millor imatge sintetitzada.

Solució proposada:

En aquest moment, l'objectiu serà definir matemàticament com aquesta distància variable *d* afecta la bondat de la imatge sintètica IM_V' i trobar el valor de *d* que la maximitza. Per cada imatge projectada IM_1', IM_2' i IM_V' definirem el seu centre: P_1, P_2 i P_V, tal com mostra la figura:

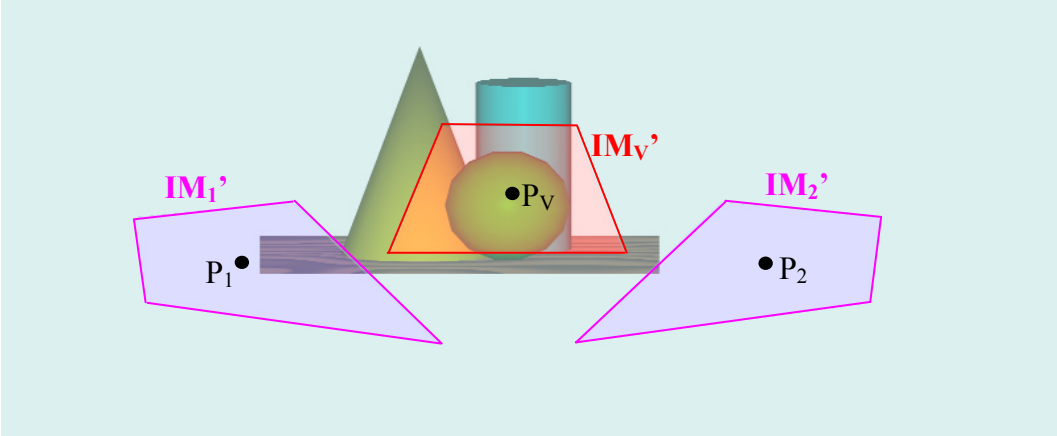


Figura 4.6 Definició per les imatges IM_1', IM_2' i IM_V' dels seus respectius centres P_1, P_2 i P_V projectats sobre el pla de reprojecció P_R.

Tots aquests valors, seran en funció del parametre *d* i s'intentara trobar el valor de *d* que els faci estar el màxim d'alineats. Considerant que s'utilitzaran línies horitzontals per la interpolació de la nova imatge, es proposa com a mesura de l'alineament entre els centres la diferència de valors de la seva component vertical, que intentarem minimitzar variant *d*. A continuació es mostra el senzill plantejament matemàtic ideat per aconseguir-ho:

1) Primer, es defineix un nou sistema de coordenades *XYZ* (ortogonal) centrat a C_1 tal com segueix:

$$X = C_2 - C_1 \quad ; \quad V = C_V - C_1 \quad ; \qquad \text{(Eq. 12)}$$

$$Z = \left(\frac{V \times X}{\| V \times X \|} \right) \cdot \| X \| \quad ; \quad Y = \left(\frac{X \times Z}{\| X \times Z \|} \right) \cdot \| X \| \quad ;$$

2) En segon lloc, es calcula la projecció del centre de la càmera virtual i de cada un dels tres vectors V_1, V_2 i V_V, damunt del pla *ZY*. Seran, respectivament, el punt $(0,Y_0)$ i els vectors V_1', V_2' i V_V' (es parla de projecció al pla *ZY*, per tant, de dues dimensions).

$$Y_0 = \frac{Y}{\| Y \|} \cdot V \quad ; \quad V_1' = \left(\frac{Z}{\| Z \|} \cdot V_1 , \frac{Y}{\| Y \|} \cdot V_1 \right) \quad ; \qquad \text{(Eq. 13)}$$

$$V_2' = \left(\frac{Z}{\| Z \|} \cdot V_2 , \frac{Y}{\| Y \|} \cdot V_2 \right) \quad ; \quad V_V' = \left(\frac{Z}{\| Z \|} \cdot V_V , \frac{Y}{\| Y \|} \cdot V_V \right) \quad ;$$

3) En tercer lloc, es defineixen les rectes r_1, r_2 i r_V com la projecció dels tres eixos òptics de les càmeres sobre el pla *ZY*.

$$r_1 \coloneqq \quad y = m_1 \cdot z \quad ; \quad m_1 = \frac{\frac{Y}{\|Y\|} \cdot V_1}{\frac{Z}{\|Z\|} \cdot V_1} \quad ;$$

$$r_2 \coloneqq \quad y = m_2 \cdot z \quad ; \quad m_2 = \frac{\frac{Y}{\|Y\|} \cdot V_2}{\frac{Z}{\|Z\|} \cdot V_2} \quad ; \qquad \text{(Eq. 14)}$$

$$r_V \coloneqq \quad y = Y_0 + m_V \cdot z \quad ; \quad m_V = \frac{\frac{Y}{\|Y\|} \cdot V_V}{\frac{Z}{\|Z\|} \cdot V_V} \quad ;$$

4) Finalment, s'ha de cercar el valor de z a les equacions de les rectes, que serà el que s'usarà després com a *d*, que minimitzi la distància vertical entre les línies. Això és la desviació definida com:

$$\sigma^2 = (m_1 \cdot d - \bar{y})^2 + (m_2 \cdot d - \bar{y})^2 + (Y_0 + m_V \cdot d - \bar{y})^2 \quad ;$$

Per a un valor promig definit per:

$$\bar{y} = \frac{m_1 \cdot d + m_2 \cdot d + Y_0 + m_V \cdot d}{3} \quad ; \qquad \text{(Eq. 15)}$$

Cercant el mínim per aquesta funció de desviació, derivant i solucionant, es troba que el valor òptim per *d* és:

$$d = Min\left(\sigma^2(d)\right) = \frac{Y_0 \cdot \left(m_1 + m_2 - 2 \cdot m_V\right)}{2 \cdot \left(m_1^{\,2} + m_2^{\,2} + m_V^{\,2} - m_1 \cdot m_2 - m_1 \cdot m_V - m_2 \cdot m_V\right)} \quad ;$$

(Solució 2)

Amb els valors de m_1, m_2 i m_V expressats a l'equació 14.

Així doncs, utilitzant aquest valor per d, es podrà sintetitzar la imatge IM_V' de mida màxima, la qual cosa millora el rendiment del mètode genèric de rectificació de tres vistes.

4.2.3 Formulació matemàtica resultant.

Amb el vist anteriorment, i assumint les premisses de que (1) es tenen les dues càmeres reals calibrades amb paràmetres intrínsecs i extrínsecs coneguts i (2) les tres càmeres estan fixades aproximadament sobre el mateix objectiu, sinó no te sentit plantejar-se la síntesi d'una nova vista**, es proposa un nou mètode** per a determinar el pla de reprojecció per la síntesi de les noves vistes. Els aspectes avantatjosos d'aquest mètode són que totes les operacions necessàries per a calcular la nova vista queden definides per uns operadors matricials que són fàcilment portables a *hardware* de cara a una implementació en temps real. Concretament, s'espera que amb l'ús d'un processador comercial d'alt rendiment, o amb la programació d'un processador digital de senyals (DSP) s'aconseguirà obtenir les imatges des d'un punt de vista variable, per suportar un ritme de *video-rate* amb una qualitat acceptable.

A continuació s'expressarà en forma d'algoritme el conjunt d'operacions necessari i que s'implementarà per comparar-lo amb altres exposats en aquesta tesi. El procediment resultant consisteix en:

1) A partir dels centres de les dues càmeres existents: C_1 i C_2, el centre de la càmera virtual C_V i els vectors que determinen els respectius eixos òptics V_1, V_2 i V_V s'aplica la solució trobada per la millor ubicació de la càmera virtual (solució 1) que permetrà calcular el centre de la càmera virtual C_S (de l'anglès *shiffted*, desplaçada).

$$C_S = C_V + \mu \cdot V_V \ . \qquad \text{(Eq. 16)}$$

Amb el valor de μ trobat a la solució 1, en el capítol 4.2.1.

2) A partir dels centres C_1, C_2 (veure figura 4.7.dalt) i el nou C_S (figura 4.7.mig) es cerca el pla P que els conté a tots tres (figura 4.7.baix).
Per trobar el pla que conté els tres centres, es parteix de les posicions dels centres, el vector de la *baseline* V_b i el vector que uneix les càmeres 1 i virtual V_{V1};

$$C_1 = \begin{bmatrix} C_{1x} \\ C_{1y} \\ C_{1z} \end{bmatrix} \quad C_2 = \begin{bmatrix} C_{2x} \\ C_{2y} \\ C_{2z} \end{bmatrix} \quad C_S = \begin{bmatrix} C_{Sx} \\ C_{Sy} \\ C_{Sz} \end{bmatrix} \quad V_b = C_2 - C_1 = \begin{bmatrix} C_{2X} - C_{1X} \\ C_{2Y} - C_{1Y} \\ C_{2Z} - C_{1Z} \end{bmatrix} \quad V_{V1} = C_V - C_1 = \begin{bmatrix} C_{VX} - C_{1X} \\ C_{VY} - C_{1Y} \\ C_{VZ} - C_{1Z} \end{bmatrix}$$

(Def. 4)

aplicant el producte vectorial dels dos vectors s'obté V_P i s'usa aquest vector per definir el pla P.

$V_P = V_{V1} \times V_b$;

Partint de l'equació del pla: $AX + BY + CZ + D = 0$;

Se'l força a passar per un dels punts (per exemple C_1) obtenint

P: $V_{PX} X + V_{PY} Y + V_{PZ} Z + D = 0$;

on $D = -(V_{PX} C_{1X} + V_{PY} C_{1Y} + V_{PZ} C_{1Z})$; (Eq. 17)

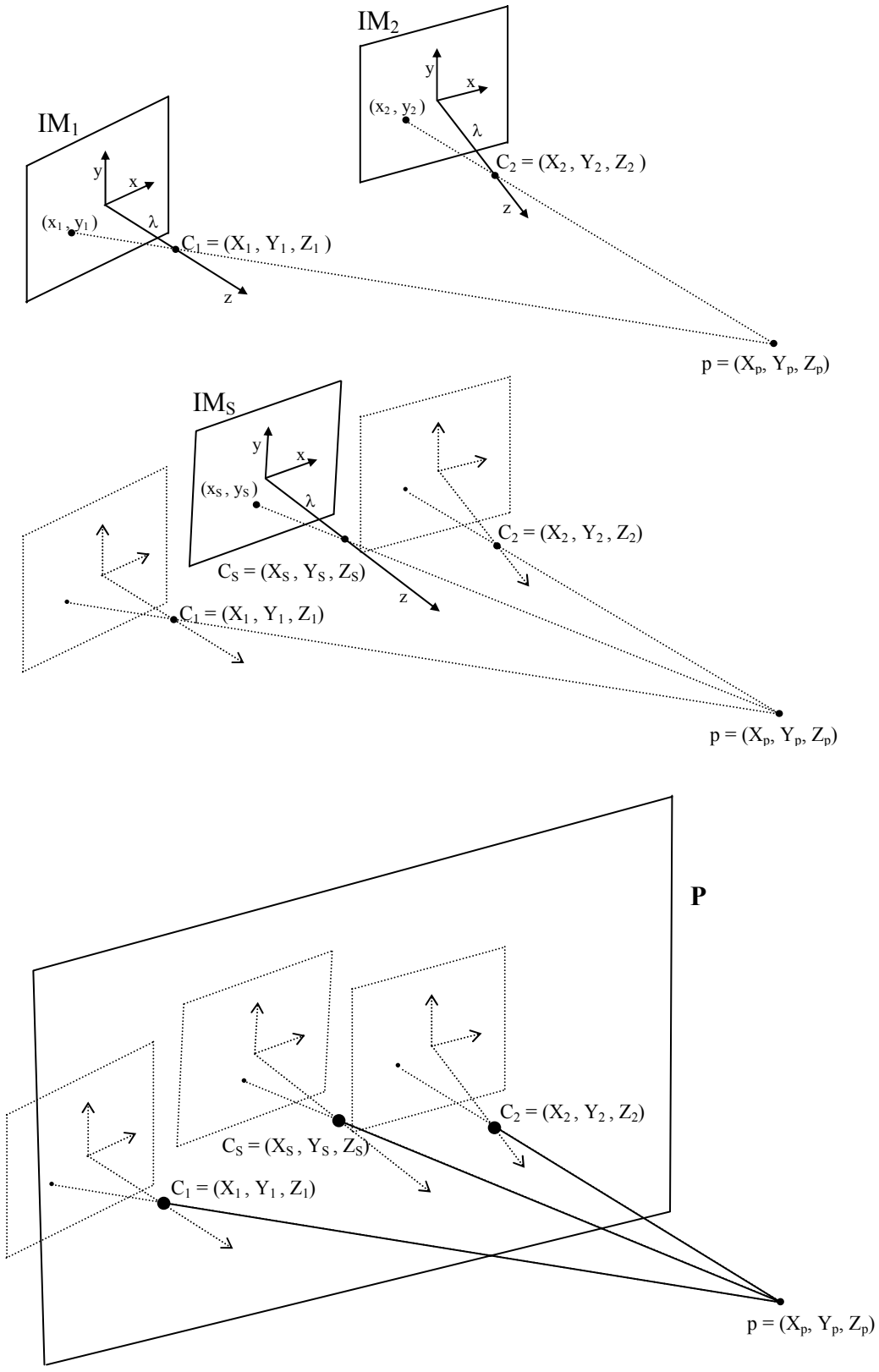


Figura 4.7 (dalt) Mostra de la ubicació de les càmeres 1 i 2 i del mecanisme de projecció d'un punt arbitrari p, (mig) mostra de la ubicació de la càmera virtual i (baix) mostra de la construcció del pla P que conté els tres centres.

3) Tot seguit, es calcula l'equació del pla de reprojecció. Partint de l'equació del pla P (equació 17), i del seu vector normal V_P, es cerca el pla paral·lel situat a una certa distància *d* del mateix.

A partir del vector normal al pla P_V:

$$V_{PV} = \begin{bmatrix} V_{PX} \\ V_{PY} \\ V_{PZ} \end{bmatrix}$$

es troba l'equació del pla paral·lel P_R: $V_{PVX}\, X + V_{PVY}\, Y + V_{PVZ}\, Z + D = 0$;

i se'l força a passar per un punt C_d situat a la distància *d* del pla P_V:

$$C_d = C_1 + d \cdot \frac{V_{PV}}{\| V_{PV} \|} \qquad \text{(Eq. 18)}$$

amb el que es D per al pla P_R : $D = -(V_{PX}\, C_{dX} + V_{PY}\, C_{dY} + V_{PZ}\, C_{dZ})$.

La figura 4.8 mostra la posició a l'espai del pla P_R.

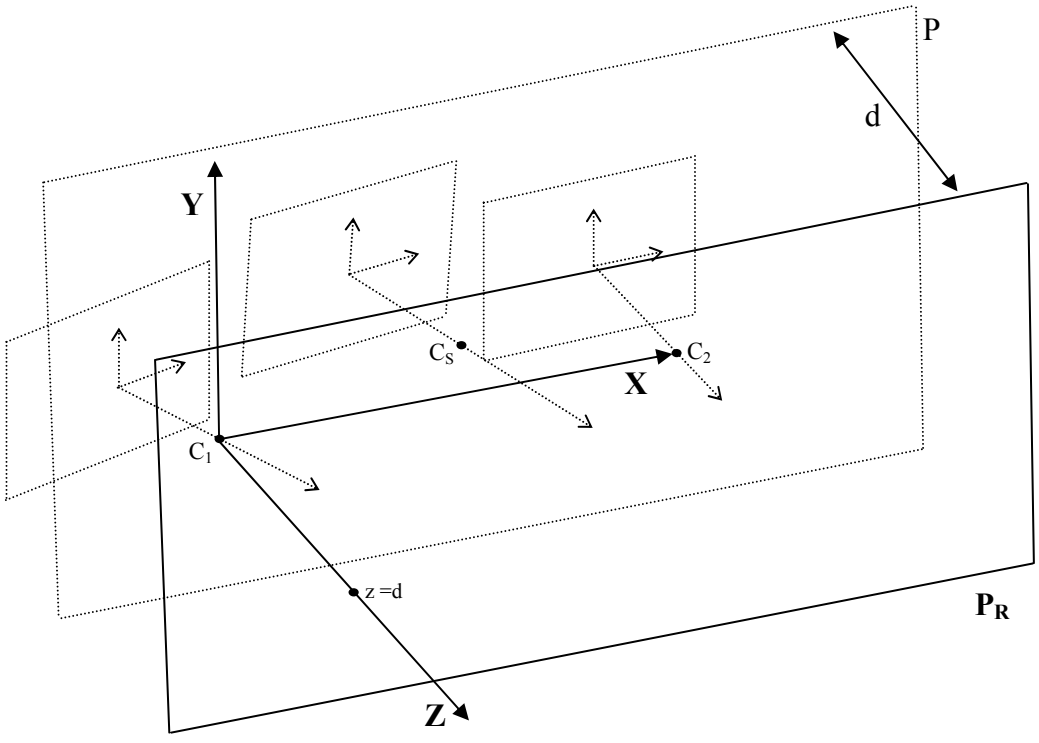


Figura 4.8 El pla de reprojecció P_R situat a distància *d* del pla P i definició del nou sistema de coordenades *XYZ*.

El valor òptim de *d* s'obté de la solució 2 exposada a l'apartat 4.2.2.

4) El següent pas a realitzar és definir un nou sistema de coordenades *XYZ* a partir de les posicions dels centres C_1, C_2 i C_S (veure figura 4.8). Aquest nou sistema de coordenades es determina en cinc passos:

   4.1) Es situa l'origen del sistema de coordenades en el punt C_1.

   4.2) Es defineix l'eix *X* damunt la recta que uneix C_1 i C_2.

   4.3) Es defineix l'eix Y com el perpendicular a l'eix X, contingut en el pla P.

   4.4) Es defineix l'eix *Z* com el perpendicular als eixos *X* i *Y*.

   4.5) Es defineix la distància unitària com la que separa C_1 i C_2 sobre l'eix *X*.

   Segons aquesta transformació, els centres de les tres càmeres C_1, C_2 i C_S tindran, en el sistema *XYZ* la següent expressió:

$$C_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad C_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad C_S = \begin{bmatrix} X_S \\ Y_S \\ 0 \end{bmatrix} \qquad \text{(Eq.19)}$$

5) Amb el nou espai *XYZ* es considera ara que les tres càmeres inicials: Càmera 1, Càmera 2 i Càmera virtual s'han convertit en tres noves càmeres amb idèntics paràmetres intrínsecs (bàsicament: distància focal i mida de la CCD), orientades de la mateixa manera, i amb els plans d'enfocament coincidents, tal com mostra la figura 4.9.

   En aquest moment, la distància focal de les tres càmeres és la distància al pla de reprojecció P_R de qualsevol dels centres: és a dir *d* i les imatges queden projectades al pla com a IM_1', IM_2' i IM_S' contingudes dins dels respectius rectangles R_1', R_2' i R_S'. Aquesta distància focal expressada en el nou sistema de coordenades *XYZ*; serà la distància *d* dividida per la distància entre C_1 i C_2, o sigui, el mòdul del vector de la *baseline*, tal com s'ha expressat en la definició 4.

$$d_{XYZ} = \frac{d}{||V_b||} \qquad \text{(Eq. 20)}$$

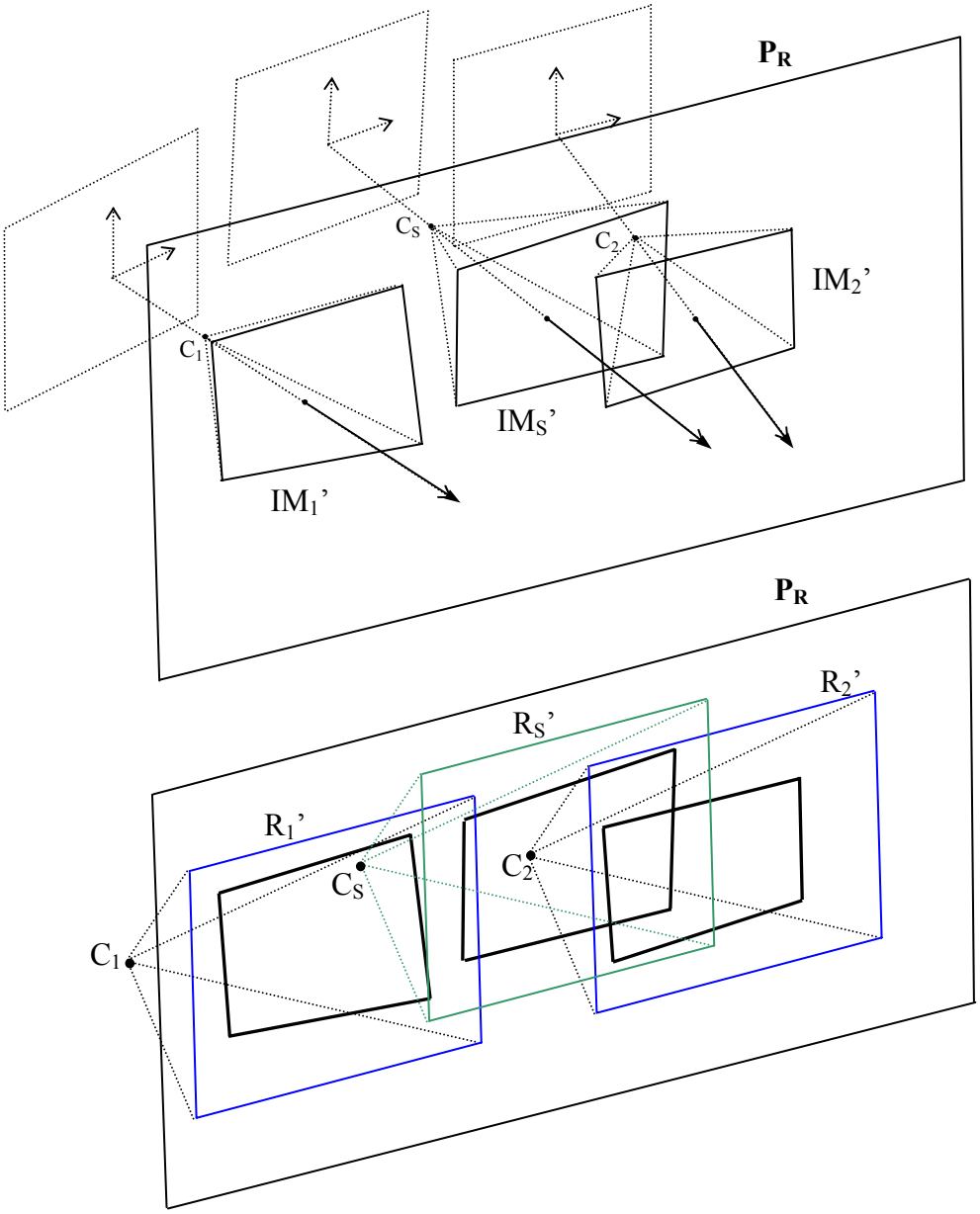


Figura 4.9 (dalt) Aspecte de les tres imatges reprojectades dins el pla P_R i (baix) aspecte de les tres noves càmeres creades amb centres C_1, C_2 i C_S i imatges R_1', R_2' i R_S' de les que es coneixen els píxels d'IM_1' i IM_2' i s'interpolaran els de IM_S'.

6) Un cop definit el nou pla de reprojecció a l'espai *XYZ* com $Z = d_{XYZ}$ en el nou sistema de coordenades; es veu que la projecció d'un punt arbitrari d'aquest espai $P = (X_P, Y_P, Z_P)$ sobre el pla $Z = d_{XYZ}$ serà respectivament el píxel p_1 per la càmera 1 sobre R_1', el píxel p_2 per la càmera 2 sobre R_2' i el píxel p_S per la càmera virtual sobre R_S' i es podran calcular amb les següents fórmules:

$$p_1 = d_{XYZ} \cdot \begin{bmatrix} X_P / Z_P \\ \\ Y_P / Z_P \end{bmatrix}; \quad p_2 = d_{XYZ} \cdot \begin{bmatrix} X_P - 1 / Z_P \\ \\ Y_P / Z_P \end{bmatrix}; \quad p_S = d_{XYZ} \cdot \begin{bmatrix} X_P - X_S / Z_P \\ \\ Y_P - Y_S / Z_P \end{bmatrix}$$

(Eq.21)

La figura 4.10 mostra el procés de projecció d'un punt de l'espai damunt d'un pla, per la coordenada X i Y separadament. Aquest és el procediment pel qual s'ha obtingut l'equació 21 aplicant la similitud de triangles pel cas X i Y.

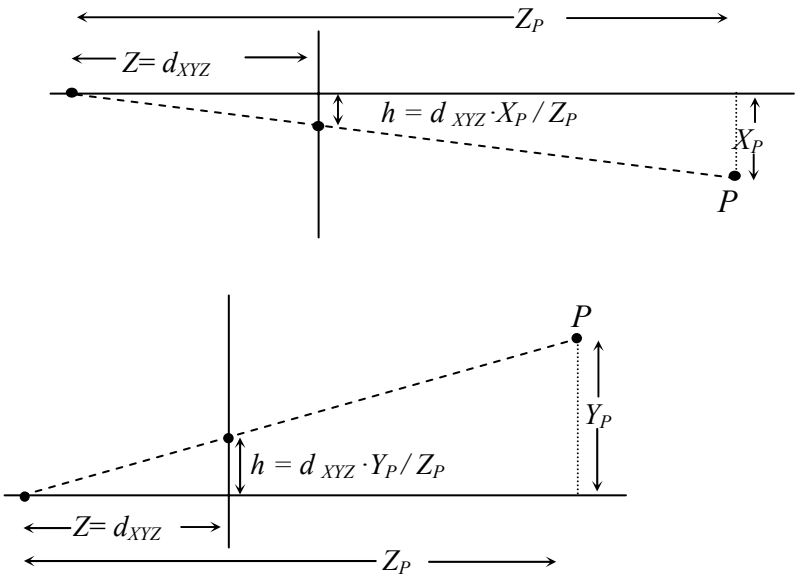


Figura 4.10 (dalt) projecció de la coordenada X d'un punt de l'espai en el pla Z= d_{XYZ} i (baix) projecció de la coordenada Y. Pel mètode de similitud de triangles podem trobar les alçades respectives que corresponen a les projeccions.

I en general, per un punt $P = (X_p, Y_p, Z_p)$ i un centre de projecció (X_{Ci}, Y_{Ci}, Z_{Ci}), el píxel corresponent en el pla $Z = d_{XYZ}$ serà (coherentment amb el vist a l'equació 21) el de coordenades:

$$p_i = d_{XYZ} \cdot \begin{bmatrix} X_P - X_{Ci} / Z_P \\ \\ Y_P - Y_{Ci} / Z_P \end{bmatrix} \qquad \text{(Eq. 22)}$$

7) Segons les premisses exposades en els punts 1 a 6 es pot calcular ara la projecció d'un punt de la imatge original IM_i sobre el pla de reprojecció $Z = d_{XYZ}$, de la mateixa manera que si fos un punt qualsevol de l'espai *P*. Fent la projecció de tots els punts de IM_i sobre el pla de reprojecció $Z = d_{XYZ}$, s'obtindrà la imatge reprojectada corresponent IM'_i. Aquest procés es repetirà per totes les imatges considerant que:

   - Com s'ha vist, per poder realitzar la reprojecció caldrà calcular les coordenades en l'espai de tots els píxels de la imatge original, i projectar-los al pla $Z = d_{XYZ}$.
   - Aquestes coordenades hauran d'estar referides en el nou sistema *XYZ*, la qual cosa permetrà aplicar la simplificació explicada a la figura 4.10 i formulada a l'equació 22.

- La projecció dels píxels de la imatge IM_i damunt de IM'_i degut a que les dues imatges tenen àrees diferents, pot portar problemes de submostreig o sobremostreig [Wolberg 90], que poden ser analitzats més endavant.

7.1) En el nou sistema de coordenades *XYZ*, un píxel p_i de la imatge original IM_i, de coordenades d'imatge (u_i, v_i), s'expressarà com:

$$p_i = O_i + u_i * R_i + v_i * S_i ; \qquad \text{(Eq. 23)}$$

On O_i és el vector origen del pla imatge IM_i respecte al centre C_i , i R_i i S_i són els vectors unitaris perpendiculars del pla imatge IM_i expressats en el nou sistema, tal com mostra la figura 4.11.

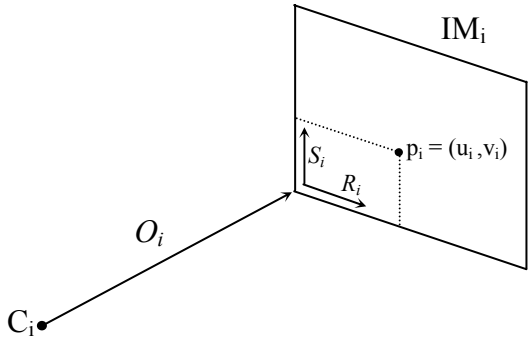


Figura 4.11. Càlcul de la posició d'un píxel p_i de la imatge original expressat segons el nou sistema de coordenades definit *XYZ*.

7.2) Com s'ha vist, per construcció del sistema (equació 22) , la projecció d'un punt $P= (X_p, Y_p, Z_p)$ sobre el pla $Z = d_{XYZ}$, sobre un dels centres de les càmeres, $C_i= (X_{Ci}, Y_{Ci}, Z_{Ci})$ correspon genèricament al píxel

$$p_i' = d_{XYZ} \cdot \begin{bmatrix} X_P - X_{Ci} / Z_P \\ \\ Y_P - Y_{Ci} / Z_P \end{bmatrix} \qquad \text{(Eq. 24)}$$

El que equival a dir que $p_i' / d_{XYZ} = P - C_i$, normalitzat.

Per normalització s'entén agafar el vector $P - C_i$ i forçar la seva tercera component a ser 1, és a dir: es resten a les coordenades del punt *P* les del centre de la càmera C_i, de les que es sap per construcció de l'espai *XYZ* que $Z_{Ci} = 0$, i es divideix el resultat de la

resta per la tercera component, per normalitzar les coordenades al pla imatge.

7.3) Combinant les expressions (Eq. 23) i (Eq. 24), fent que el punt P que es projecta al píxel p_i' de la imatge reprojectada sigui el píxel p_i de la imatge original, es té que:

$$p_i' / d_{XYZ} = p_i - C_i = u_i * R_i + v_i * S_i + O_i - C_i \qquad \text{(Eq. 25)}$$

Que es pot expressar de forma matricial com:

$$p_i' / d_{XYZ} = \left[\begin{array}{c|c|c} R_i & S_i & O_i - C_i \end{array}\right] \times \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} \qquad \text{(Eq. 26)}$$

7.4) Finalment, es defineix que, la matriu que relaciona els píxels de la imatge original expressats com $p_i = (u_i, v_i, 1)$ amb els píxels en la imatge reprojectada p_i' s'anomena matriu d'homografia i permetrà passar ràpidament d'una imatge a l'altra (cal recordar que d_{XYZ} és tan sols un valor escalar).

Determinant les matrius $\mathbf{H_1}$ i $\mathbf{H_2}$ es podran trobar les dues imatges originals IM_1 i IM_2 projectades en el pla $Z = d_{XYZ}$. En aquest pla es farà en condicions epipolars la síntesi de la nova vista, que es projectarà a la càmera virtual utilitzant la matriu inversa $\mathbf{H_S^{-1}}$. Les tres matrius que s'usaran en els càlculs seran:

$$H_1 = \left[\begin{array}{c|c|c} R_1 & S_1 & O_1 - C_1 \end{array}\right]; \qquad H_2 = \left[\begin{array}{c|c|c} R_2 & S_2 & O_2 - C_2 \end{array}\right];$$

$$H_S^{-1} = \left[\begin{array}{c|c|c} R_S & S_S & O_S - C_S' \end{array}\right]^{-1} \qquad \text{(Eq. 27)}$$

On els vectors R_1, R_2, R_S, S_1, S_2 i S_S, i les coordenades dels punts O_1, O_2, O_S, C_1, C_2 i C_S, s'hauran d'expressar en el sistema de coordenades *XYZ* definit a l'apartat 4 d'aquest subcapítol. Per això es podrà utilitzar una matriu de canvi de sistema de coordenades que s'anomenarà $\mathbf{M_{XYZ}}$.

8) Una vegada obtinguda la imatge des del punt de vista C_S amb l'aplicació de la matriu inversa d'homografia, caldrà obtenir la vista corresponent al punt C_V (que és el desitjat inicialment). Per disposar de la vista des de C_V caldrà desfer el canvi efectuat a l'apartat 1 d'aquest subcapítol i això es farà mitjançant una matriu d'escalat $\mathbf{M_S}$. La figura 4.12 mostra el significat físic de la matriu d'escalat necessitada.

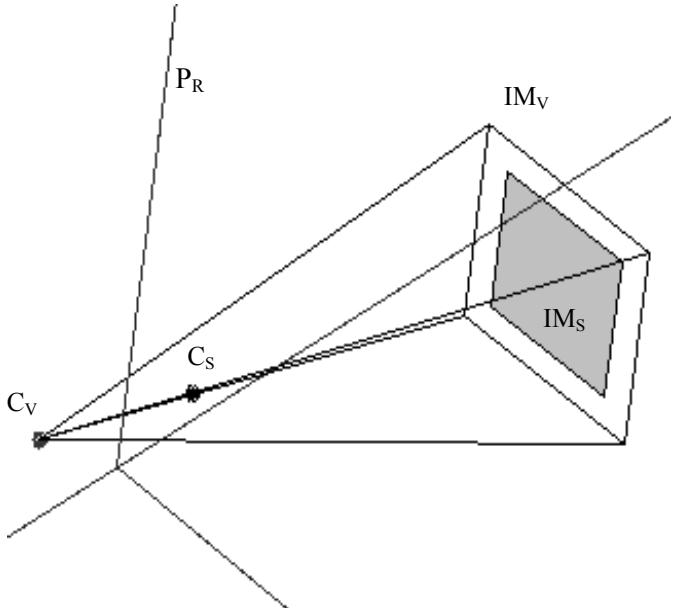


Figura 4.12. Relació entre la imatge projectada des de el punt de vista virtual original C_V i el desplaçat C_S.

9) **Ús dels mapes de disparitat.**

Un cop obtingudes les imatges reprojectades IM_1' i IM_2' a través de les matrius d'homografia caldrà tenir els mapes de disparitat entre les dues imatges; que s'anomenaran D_{12} i D_{21}.
Cada posició en el mapa de disparitat D_{12}, donarà la distància del píxel (x, y) de la imatge IM_1' al seu corresponent en la imatge IM_2' sobre el pla de reprojecció. Per això, caldrà tenir un algoritme d'aparellament que realitzi, per a cada punt, la cerca del seu punt corresponent en una zona determinada de la imatge IM2'. Gràcies a la configuració obtinguda amb la reprojecció i el canvi de sistema de coordenades obtingut, el punt corresponent caldrà cercar-lo només damunt un segment rectilini (configuració epipolar).
De moment, es suposa que l'algoritme existeix i funciona, i que ja es disposa dels mapes de disparitat. En el capítol cinquè s'analitzarà la problemàtica concreta dels algoritmes d'aparellament estèreo i altres vies per obtenir ràpidament el mapa de disparitat aplicats a la generació de noves vistes, els seus requeriments i les aportacions necessàries per a la seva optimització.

9.1) Si suposem que p_1 i p_2 són les projeccions d'un punt de l'espai P en les imatges IM_1' i IM_2' respectivament, per construcció del sistema (veure

figura 4.13), ambdues projeccions mesuraran la següent distància entre elles (a partir de l'equació 4.22):

$$p_i = d_{XYZ} \cdot \begin{bmatrix} X_P - X_{Ci} / Z_P \\ \\ Y_P - Y_{Ci} / Z_P \end{bmatrix}$$

$$d_{12} = \left[\ p_2 - p_1\ \right]_X = d_{XYZ} \cdot \left[\frac{X_P - 1}{Z_P} - \frac{X_P - 0}{Z_P} \right] = d_{XYZ} \cdot (-1 / Z_P)$$

$$d_{21} = \left[\ p_1 - p_2\ \right]_X = d_{XYZ} \cdot \left[\frac{X_P - 0}{Z_P} - \frac{X_P - 1}{Z_P} \right] = d_{XYZ} \cdot (1 / Z_P)$$

(Eq. 28)

9.2) La component Y de la distància entre píxels serà 0 degut a la rectificació.

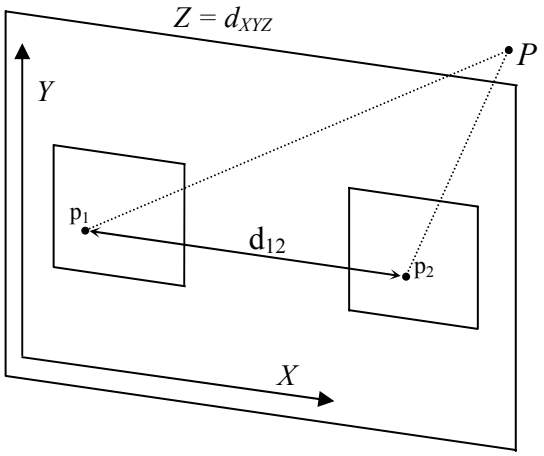


Figura 4.13. Definició de la disparitat entre els píxels corresponents a la projecció d'un punt P en dues vistes diferents.

9.3) Es calcularan ara les coordenades del corresponent punt a la imatge virtual IM_S. S'anomenarà p_S a la projecció del punt *P* en el pla $Z = d_{XYZ}$, dins de la imatge virtual desplaçada, i de la qual es volem conèixer les coordenades. El primer que es farà serà, a partir de la definició de les projeccions, mostrada en l'equació 28, trobar la distància entre p_S i les projeccions p_1 i p_2. El valor d'aquestes distàncies és:

$$p_S - p_1 = d_{XYZ} \cdot \begin{bmatrix} -X_S / Z_P \\ -Y_S / Z_P \end{bmatrix}; \quad p_S - p_2 = d_{XYZ} \cdot \begin{bmatrix} -(X_S - 1) / Z_P \\ -Y_S / Z_P \end{bmatrix}$$

(Eq. 29)

Aïllant de les equacions p_S i tenint en compte els valors de la distància entre p_1 i p_2, i p_2 i p_1 (veure equació 28), es pot deduir l'anomenada equació de *warping* lineal [Scharstein 99] que permetrà, a partir dels mapes de disparitat i de la posició del centre de la càmera virtual, obtenir ràpidament les coordenades dels píxels de la imatge sintetitzada:

$$p_S = p_1 + d_{12} \cdot \begin{bmatrix} X_S \\ Y_S \end{bmatrix}; \qquad p_S = p_2 - d_{21} \cdot \begin{bmatrix} X_S - 1 \\ Y_S \end{bmatrix} \qquad \text{(Eq. 30)}$$

Així doncs, qualsevol de les formulacions de l'equació 30, es prendrà la primera per ser més senzilla, permetrà obtenir a partir dels punts de la imatge original rectificada, i la informació dels mapes de disparitat, la imatge sintetitzada completa. Sols faltarà aplicar la matriu inversa d'homografia H_S^{-1} per trobar la imatge vista des del nou centre virtual C_S i finalment, aplicar un escalat sobre la imatge per recuperar el punt de vista desitjat C_V.

4.3. Descripció algorísmica del mètode millorat.

L'apartat 4.2.3 ha descrit en detall el mètode de rectificació de tres vistes modificat per obtenir un funcionament òptim. Els punts 1 a 9 han anat mostrant els passos a seguir des del moment inicial en que es coneixen els punts de vista reals i virtual i es tenen només les imatges originals, fins al final, on s'obté la imatge sintetitzada. En aquest darrer apartat del capítol dedicat a la síntesi de vistes es mostra la descripció algorísmica utlitzada per facilitar-ne la implementació amb un processador dedicat com és ara un DSP. Els experiments realitzats amb el mètode de síntesi han demostrat que, en un computador actual, el temps d'execució és d'entre 200 i 300ms, això serà insuficient per aplicacions interactives, i per això es planteja aquesta implementació alternativa.

A continuació es mostrarà, amb metodologia de disseny descendent, l'especificació complerta de l'algoritme. En els diagrames mostrats, s'ha emprat la següent codificació:

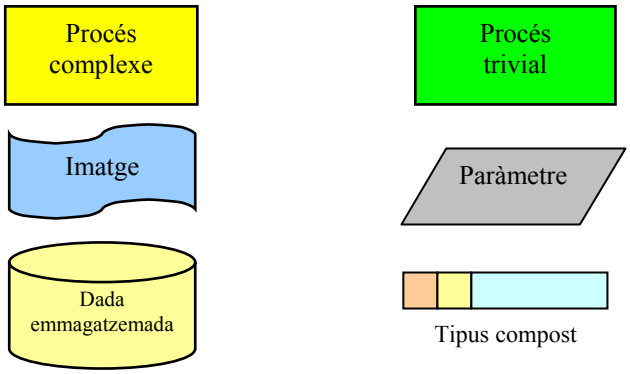


Figura 4.14 Llegenda per la comprensió dels diagrames de blocs de l'especificació de l'algoritme.

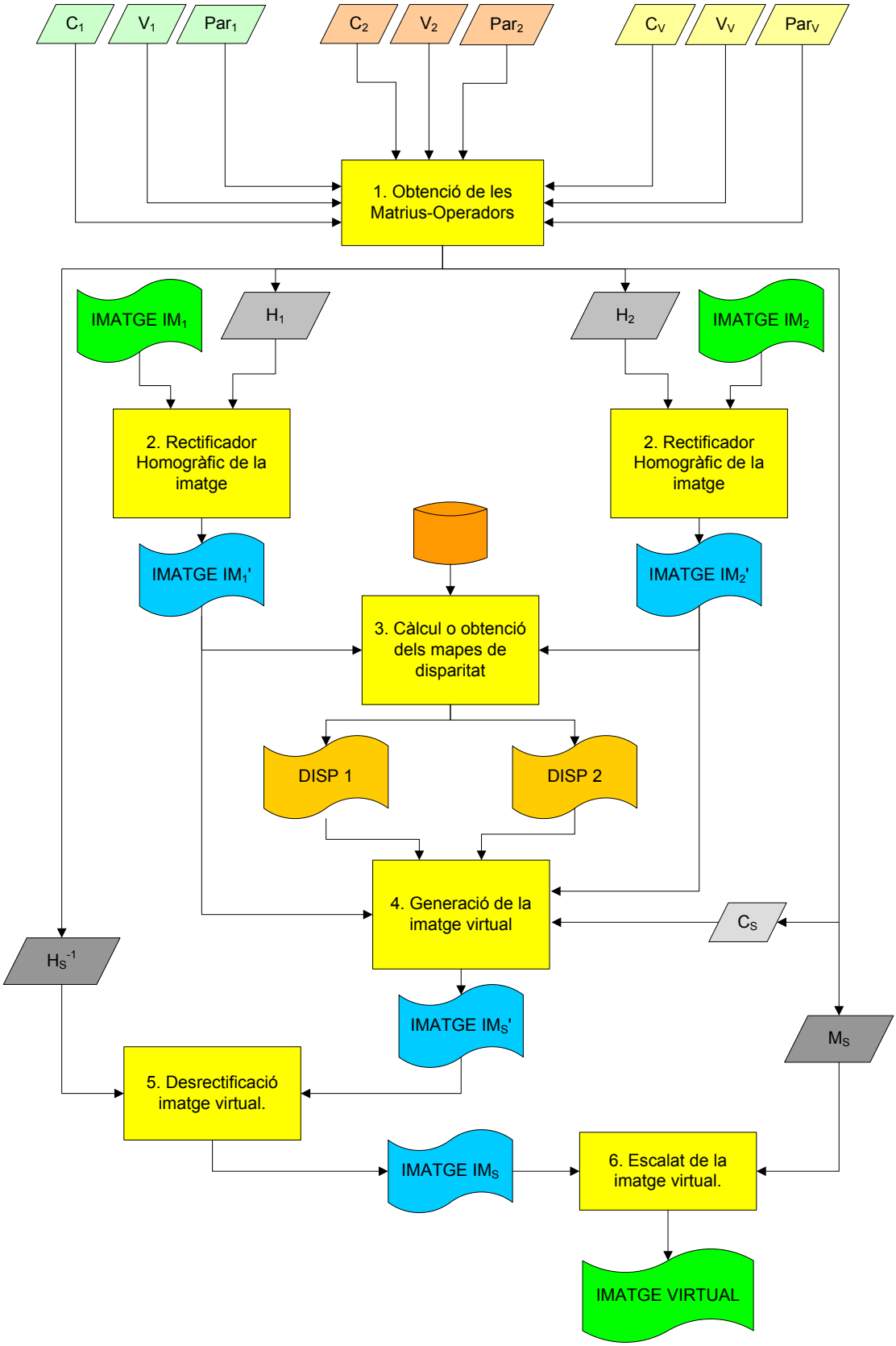
S'inicia aquesta especificació en aquest apartat, amb el diagrama de blocs general del procés.

Esquema general: Es pot considerar el mètode de síntesi de vistes com una F funció:

$$F\,(C_1, C_2, C_V, V_1, V_2, V_V, PAR_1, PAR_2, PAR_V, IM_1, IM_2) \rightarrow IM_V$$

On:

C_i, V_i, PAR_i : Centre, vector eix òptic i paràmetres de la càmera i.
IM_1, IM_2, IM_V : Imatges reals i virtual

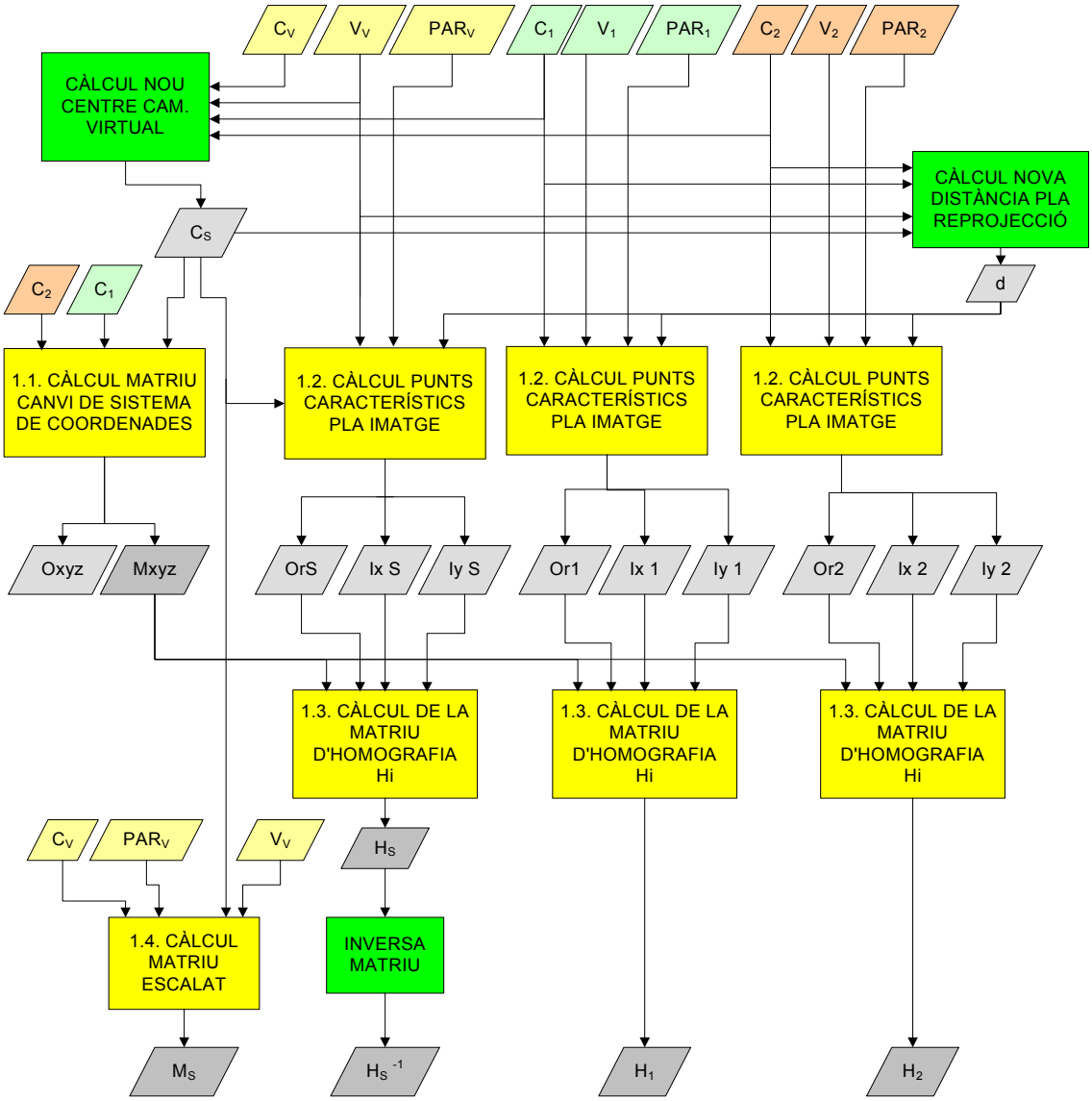


4.3.1 Obtenció de les matrius operadors.

$F\,(C_1, C_2, C_V, V_1, V_2, V_V, PAR_1, PAR_2, PAR_V) \rightarrow H_1, H_2, H_S^{-1}, M_S$

On:

C_1, V_1, PAR_1 : Centre, vector i paràmetres de la càmera 1.
C_2, V_2, PAR_2 : Centre, vector i paràmetres de la càmera 2.
C_V, V_V, PAR_V : Centre, vector i paràmetres de la càmera virtual.
H_1, H_2 : Matrius de transformació homogràfica per les càmeres 1 i 2.
H_S^{-1}: Matriu de transformació homogràfica inversa per la càmera virtual.
M_S : Matriu d'escalat per la càmera virtual.



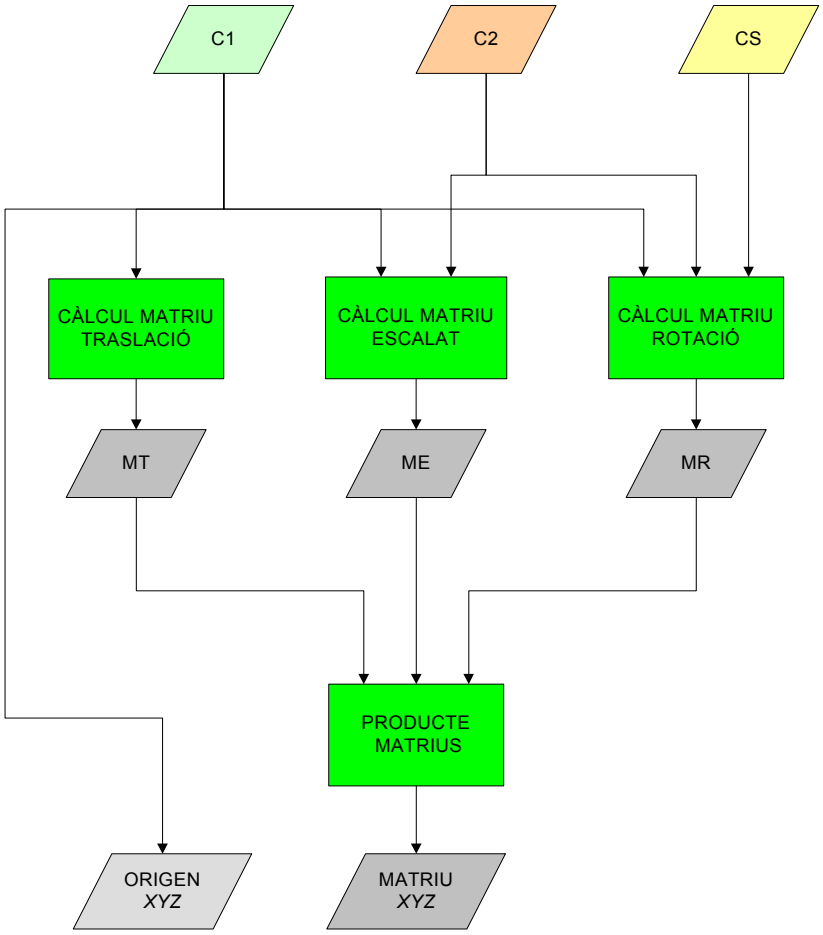
L'objectiu d'aquest primer procés és el d'obtenir les matrius d'homografia per rectificar les imatges i la matriu d'escalat.

4.3.1.1 Càlcul de la matriu de canvi de sistema de coordenades M_{XYZ} i del nou origen de coordenades.

$F\ (C_1, C_2, C_S) \rightarrow M_{XYZ}\ ,\ O_{XYZ}$.

On:

C_1, C_2, C_S: Centres de les càmeres existents i nou centre de la càmera virtual.
M_{XYZ} : Matriu de transformació de sistema de coordenades.
O_{XYZ} : Origen del nou sistema de coordenades.



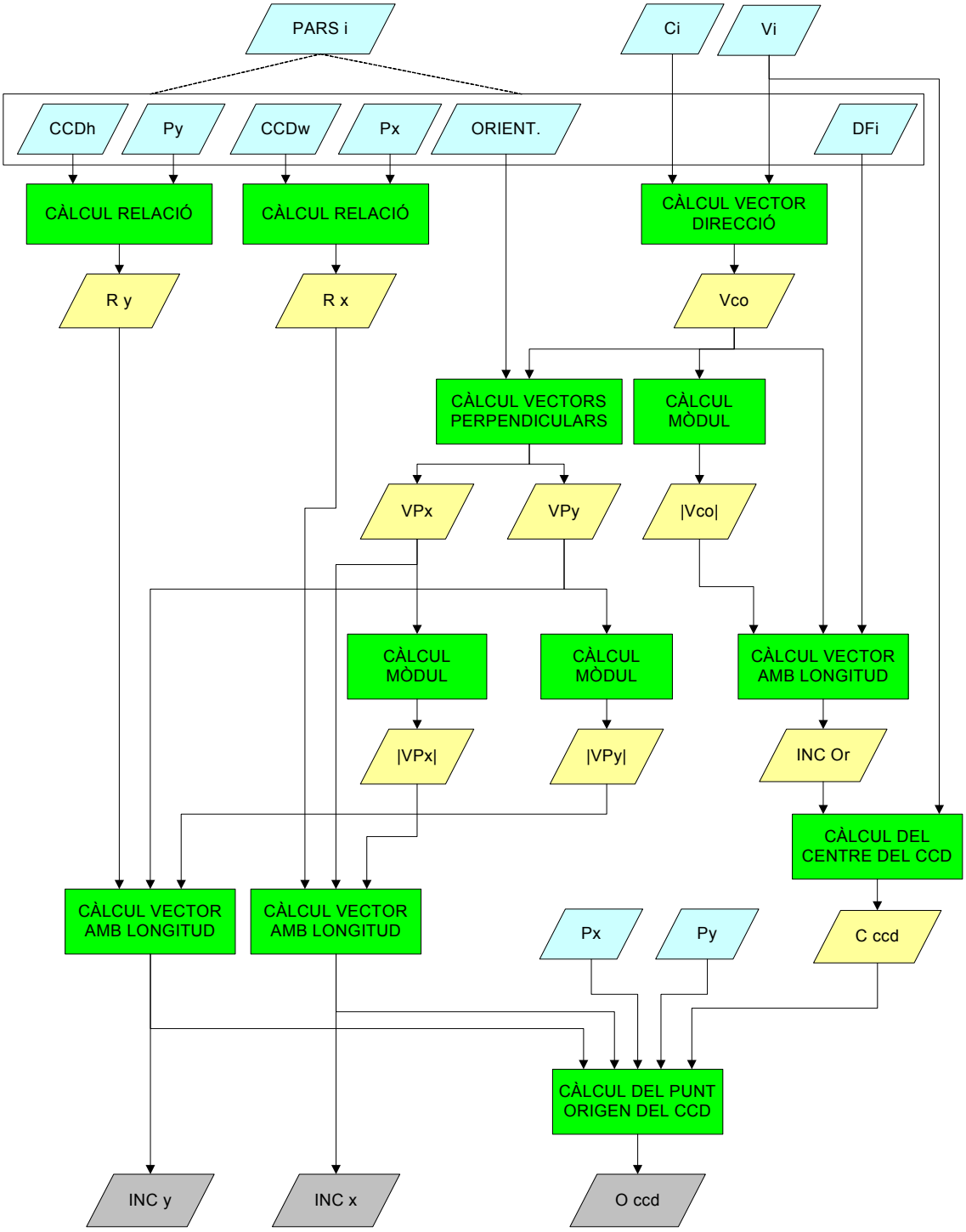
Aquest procés calcula la matriu de canvi de sistema de coordenades del definit inicialment al construït pels centres de les tres càmeres segons s'ha explicat en el quart pas de l'apartat 4.3. El resultat és la matriu de canvi de sistema i el punt origen del nou sistema de coordenades.

4.3.1.2 Càlcul dels punts característics del pla imatge.

F (Ci, Vi, DFi, CCDWi, CCDHi, PX, PY) → *OCCD, INCX, INCY*

On:

Ci ,Vi i DFi : Centre, vector eix òptic i distància focal de la càmera en qüestió.
CCDWi, CCDHi : Dimensions del CCD.
PX, PY: Dimensions en píxels de la imatge.
OCCD: Punt origen del pla imatge
INCX, INCY: Increments X i Y entre els punts projectats.



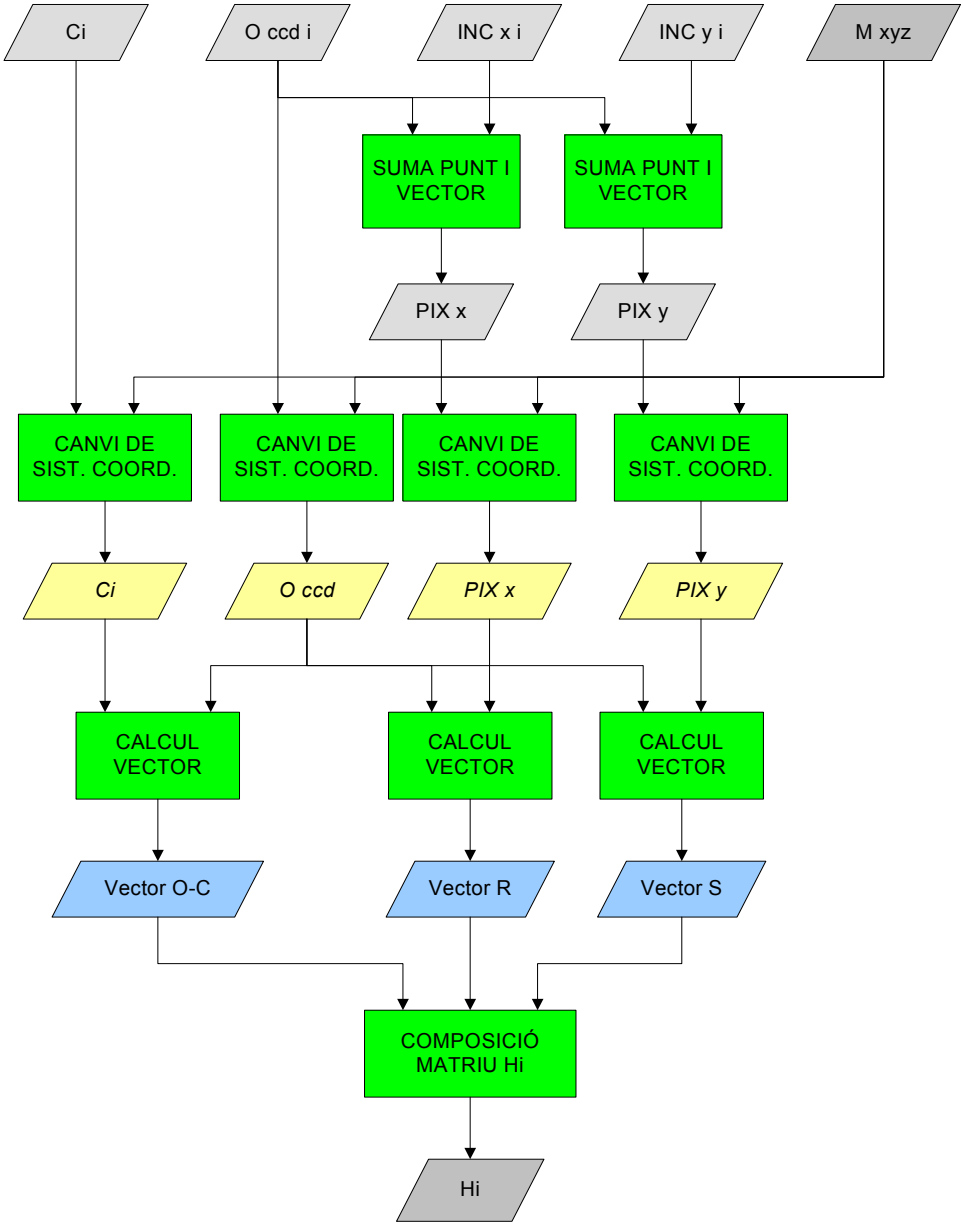
Aquest procés calcula per cada càmera, l'expressió de l'origen de la seva CCD i dels vectors X i Y del pla CCD en el nou sistema de coordenades.

4.3.1.3 Càlcul de la matriu d'homografia Hi (per la càmera i-èssima)

F (Ci, OCCD i ,INCX i, INCY i, MXYZ) → Hi

On:

Ci: Centre de la càmera en qüestió.
OCCD i: Punt origen del pla imatge.
INCX i, INCY i: Increments X i Y entre els punts projectats.
MXYZ i : Matriu de transformació de sistema de coordenades.
Hi: Matriu de transformació d'homografia per la càmera i.



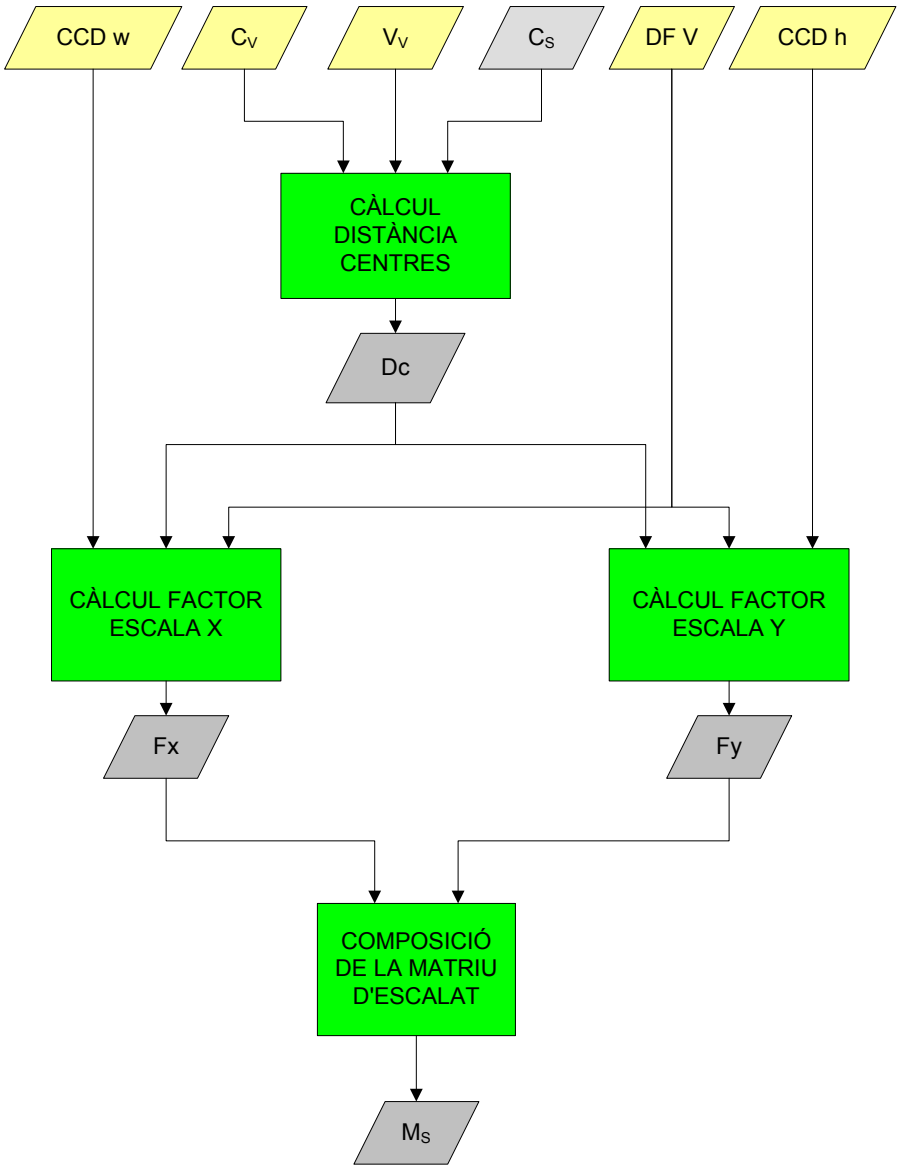
Aquest procés calcula, per cada una de les càmeres, la matriu d'homografia que permetrà passar un píxel del sistema de coordenades original, a un píxel en la CCD rectificada i referit en el nou sistema de coordenades.

4.3.1.4 Càlcul de la matriu d'escalat.

$F\,(C_V, V_V, C_S, CCDW, CCDH, DFV) \rightarrow M_S$

On:

C_V: Centres de la càmera virtual original.
V_V : Vector de la càmera virtual.
C_S: Nou centre proposat per la càmera virtual.
CCDW ,CCDH ,DFV: Paràmetres de la càmera virtual; mides del CCD i distància focal.
ME: Matriu d'escalat per passar de la imatge del nou punt de vista virtual a l'original.



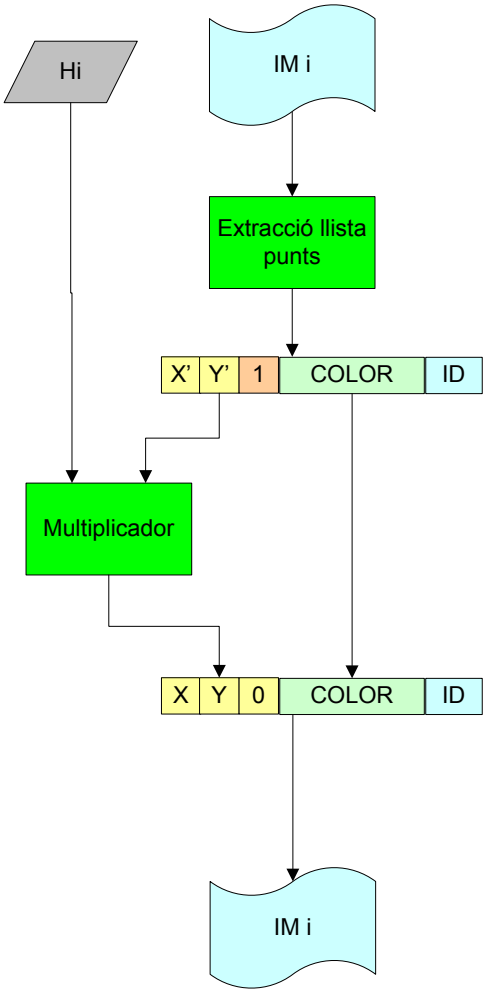
Aquest procés calcula, en cas de que el centre de la càmera virtual hagi estat modificat per evitar el problema de la impossibilitat de rectificació de les imatges, una matriu d'escalat de la imatge virtual, que permeti compensar aquest moviment endavant o endarrera de la càmera (tal com s'ha definit en l'apartat 4.2.1)

4.3.2 Rectificador homogràfic de la imatge.

$F\ (IM_i, H_i) \rightarrow IM_i'$

On:

IM_i : Imatge obtinguda per la càmera i.
H_i: Matriu de transformació homogràfica per les càmeres i.
IM_i' : Imatge i rectificada.



Aquest és el procés que aplica la rectificació a cada una de les imatges per expressar-les en el nou sistema de coordenades, on serà més senzill interpolar la nova vista. L'expressió dels píxels en forma de vector permet la multiplicació per la matriu d'homografia. Cada element inclourà les seves coordenades, el seu color i per si la informació tridimensional (disparitat) ha estat calculat prèviament, un identificador per trobar instantàniament el seu punt corresponent a l'altra vista rectificada.

4.3.3 Càlcul o obtenció del mapa de disparitat.

$F\ (IM_1', IM_2') \rightarrow DISP_1, DISP_2$

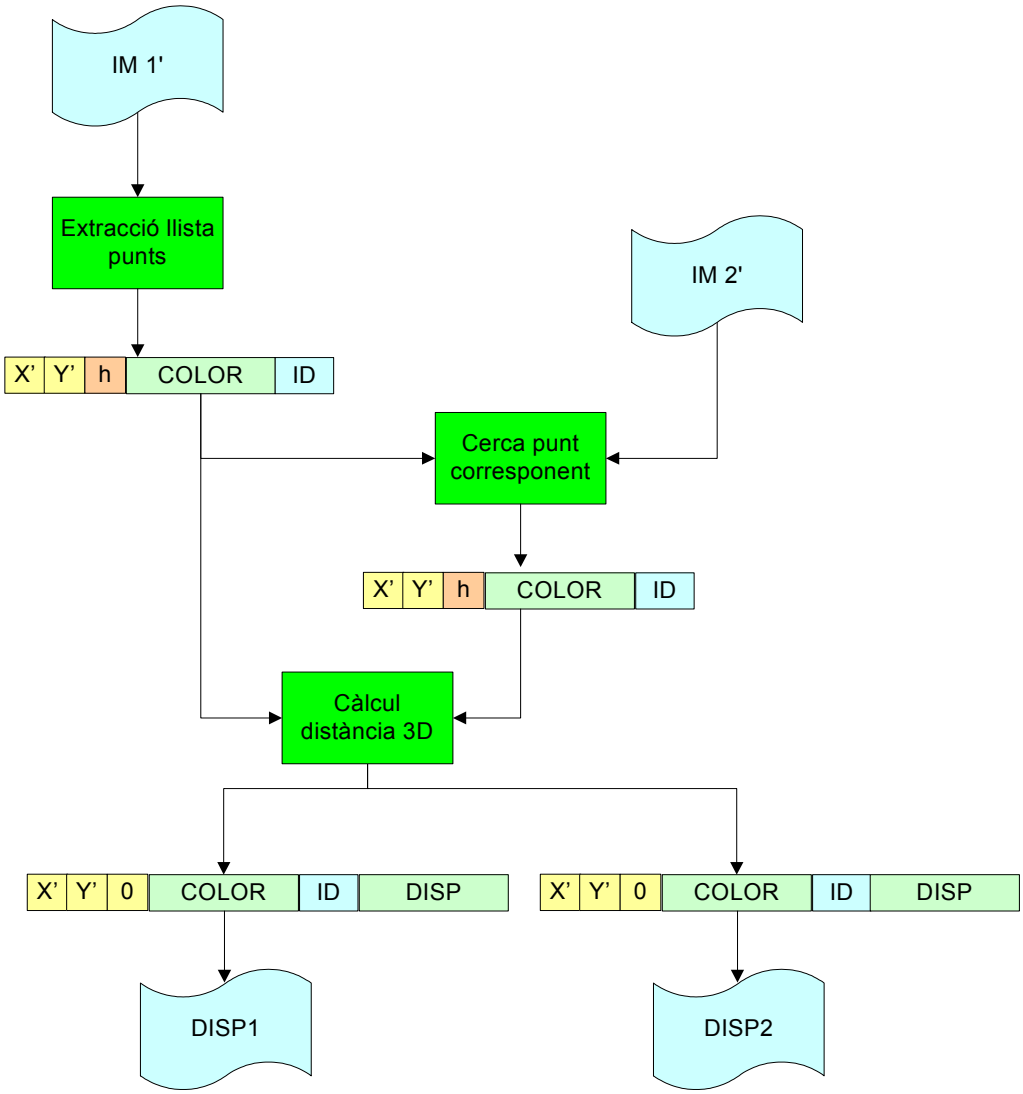
On:

IM_1' : Imatge i rectificada.

IM_2' : Imatge i rectificada.

$DISP_1$: Mapa de disparitat imatge 1 a 2.

$DISP_2$: Mapa de disparitat imatge 2 a 1.



S'expressa aquí la primera possibilitat per calcular el mapa de disparitat, consistent en, cercar amb un algoritme de correspondència estèreo, la projecció del píxel a l'altra imatge. Existirà la versió alternativa, més fàcil per a un procés interactiu, de precalcular les correspondències i arrossegar identificadors dels punts corresponents, com s'expressa a continuació:

$F\ (IM_1', IM_2', 3D) \rightarrow DISP_1, DISP_2$

On:

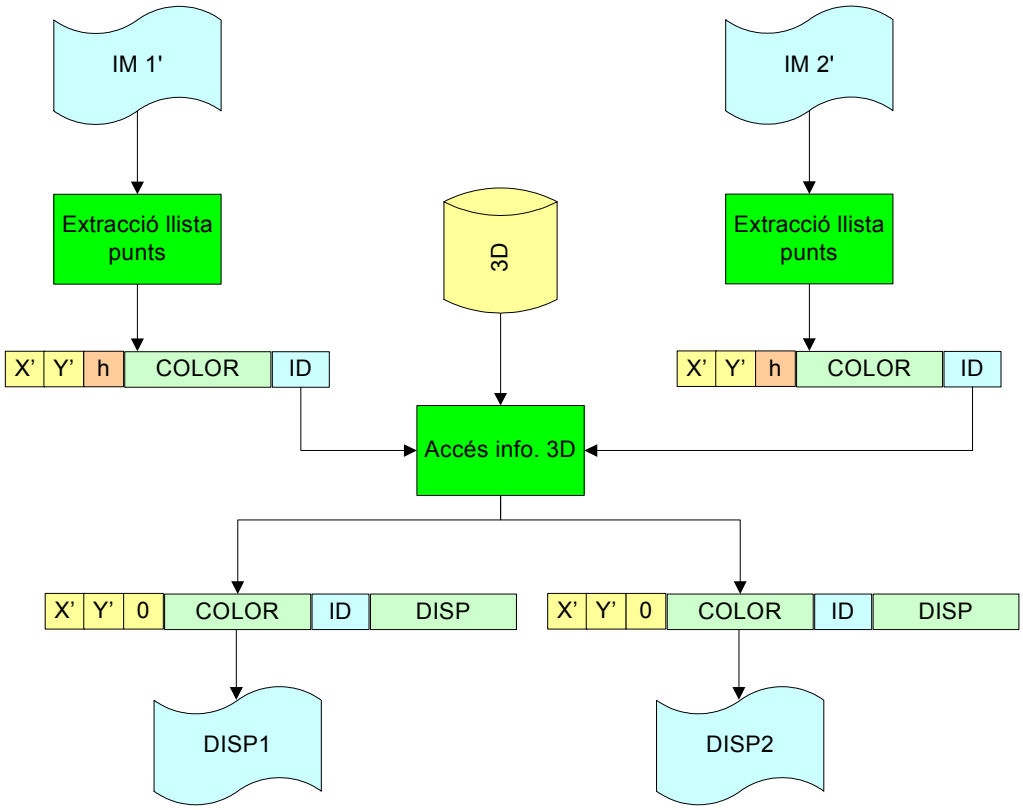
IM_1' : Imatge i rectificada.

IM_2' : Imatge i rectificada.

3D: informació tridimensional obtinguda prèviament.

$DISP_1$: Mapa de disparitat imatge 1 a 2.

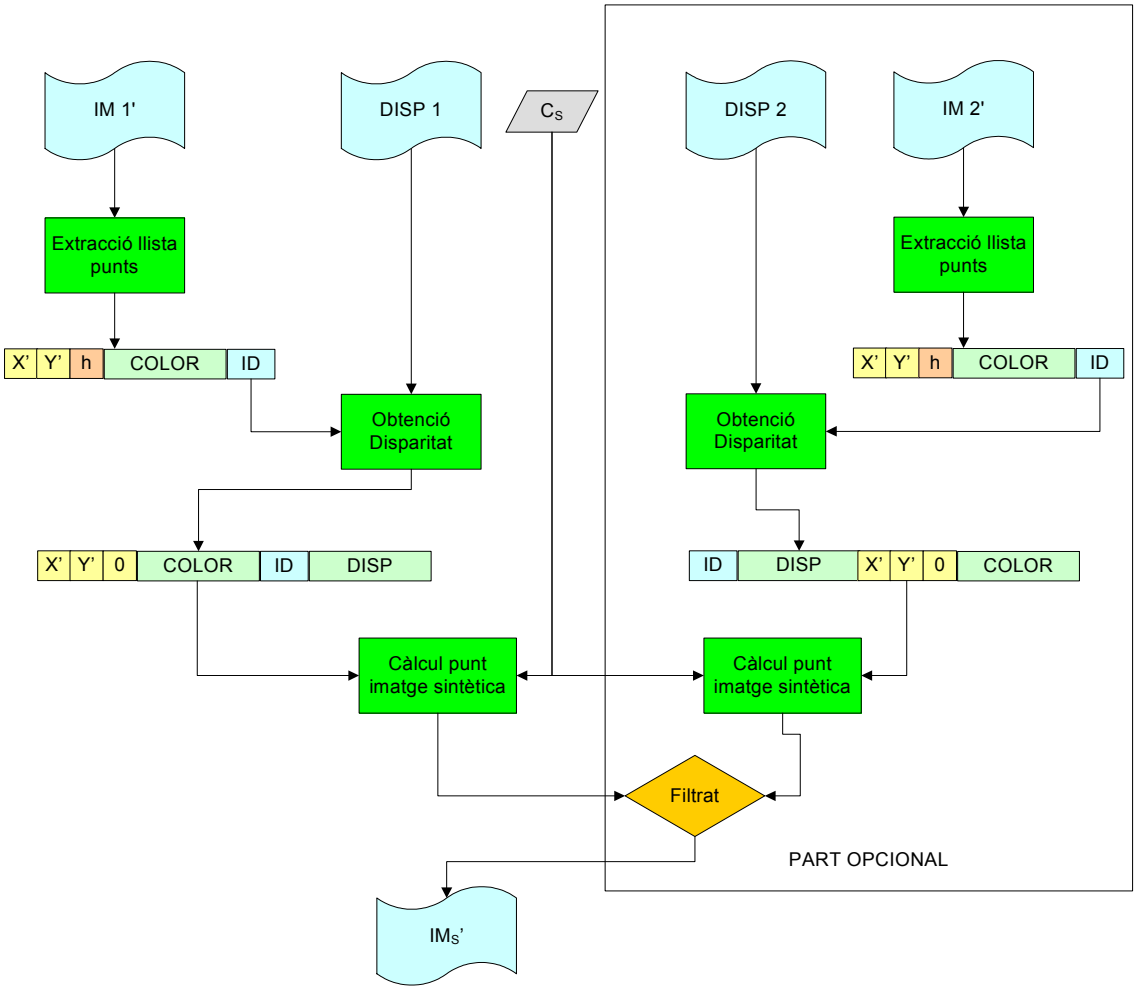
$DISP_2$: Mapa de disparitat imatge 2 a 1.



En aquesta segona proposta d'implementació, que es preferirà per implementacions interactives, s'haurà adjuntat a les vistes informació de correspondència entre els punts de les vistes esquerra i dreta. En el capítol cinquè es mostra com s'ha obtingut aquesta informació.

4.3.4 Generació de la imatge virtual.

F (IM_1', IM_2', C_S, $DISP_1$,$DISP_2$) → IM_S'
On:
IM_1' : Imatge i rectificada.
IM_2' : Imatge i rectificada.
C_S: Nou centre de la càmera virtual.
$DISP_1$: Mapa de disparitat imatge 1 a 2.
$DISP_2$: Mapa de disparitat imatge 2 a 1.
IM_S' : Imatge virtual rectificada.



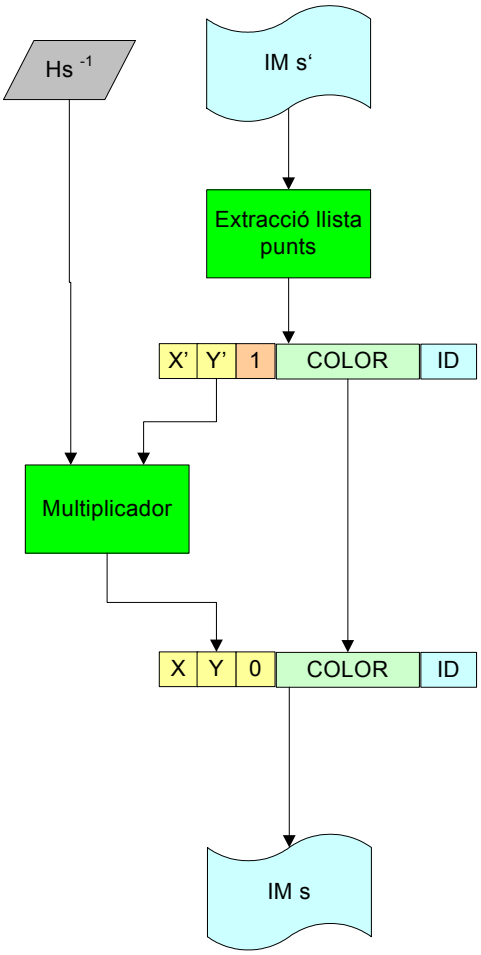
S'entra ara a la part final del mètode; a partir de les vistes i la informació de disparitat, s'obté la imatge sintètica rectificada. Per tots els punts de les vistes originals, es calculen les projeccions en la vista virtual, omplint la nova imatge. Un procés de filtrat d'errors i/o interpolació, permetrà millorar la qualitat (només visual) de la imatge resultant.

4.3.5 Desrectificació de la imatge virtual.

$F\ (IM_S', H_S^{-1}) \rightarrow IM_S$

On:

IM_S' : Imatge virtual rectificada.
H_S^{-1} : Imatge i rectificada.
IM_S : Imatge virtual desrectificada.

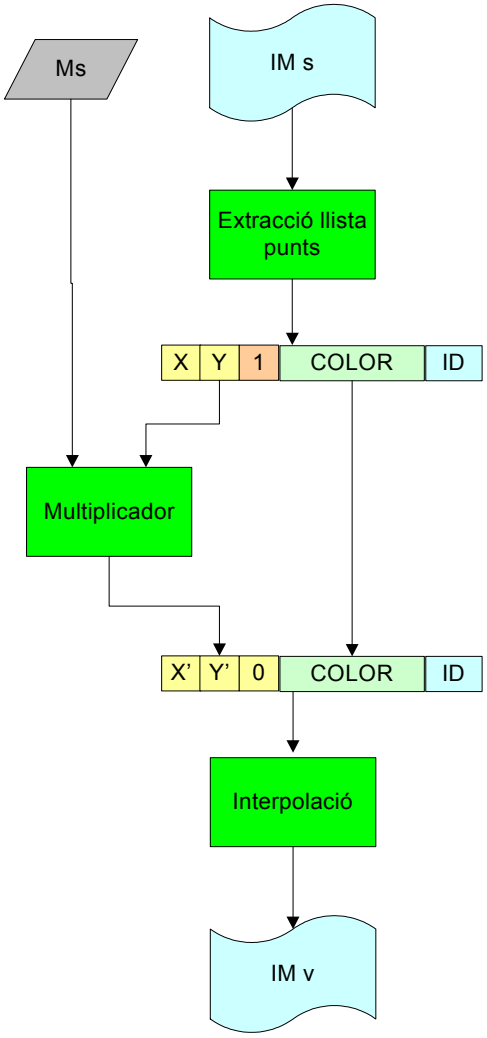


El penúltim pas del procés serà agafar la llista de punts de la imatge virtual sintetitzada en el nou sistema de coordenades i traslladar-lo a les coordenades originals de la CCD on hi ha la càmera virtual. Aquest procés s'ha anomenat desrectificació i també consistirà en l'aplicació seqüencial del producte d'una matriu per la llista de vectors de punts de la imatge.

4.3.6 Escalat de la imatge virtual.

$F\ (IM_S\ ,M_S) \rightarrow IM_V$
On:

IM_S' : Imatge virtual rectificada.
M_S : Matriu escalat.
IM_V : Imatge virtual.



Finalment, si s'havia produït el moviment de la càmera virtual, caldrà corregir-lo amb l'aplicació de la matriu d'escalat sobre la llista de punts de la imatge, obtenint la nova vista.

5. Obtenció dels mapes de disparitat. Reconstrucció 3D.

En el capítol quart s'ha vist que per a poder sintetitzar noves vistes a partir de dues ja obtingudes és necessari tenir informació de disparitat entre els seus píxels. Aquesta informació normalment es mostra en forma d'imatge on, el nivell de gris representa la distància en píxels d'un punt donat de la primera imatge al seu corresponent de la segona (veure figura 5.1). Existeix una relació entre aquest valor i la profunditat del punt observat per aquell píxel; concretament la disparitat és inversament proporcional a la distància del punt observat a les càmeres.

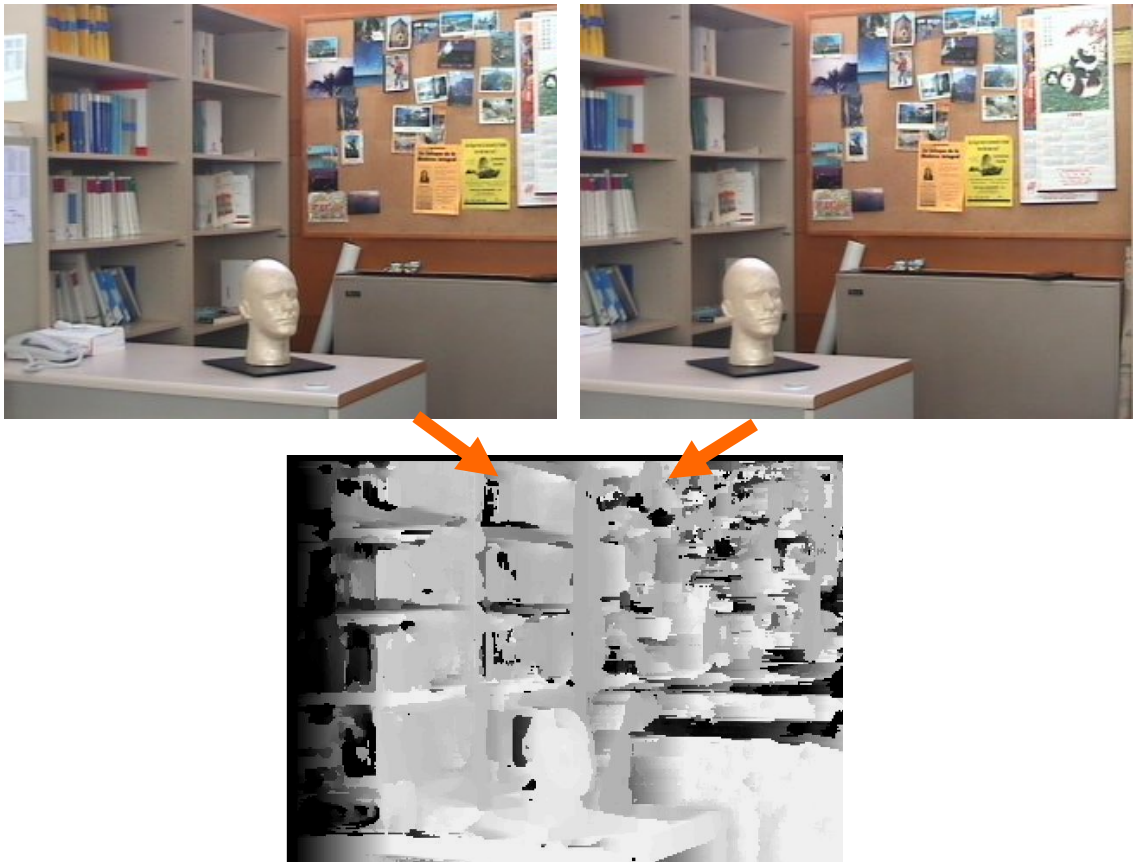


Figura 5.1 Mapa de disparitat obtingut amb un algoritme SSD i interpolació amb una finestra de 7 píxels. El seu temps de còmput (per parell d'imatges en PC Pentium IV 3GHz) és de 32 segons.

Com es veurà la informació tridimensional de l'escena pot tractar-se de diferents maneres: la més explícita és dibuixar un model 3D de la mateixa amb alguna eina de dibuix o *render*, la menys és mostrar un mapa de disparitat com a imatge, pel camí queden estructures com els mapes de correspondència entre píxels on es guarda una llista de vincles entre parells de punts. En aquest capítol es veurà com obtenir aquesta informació per diferents vies i com es poden usar per generar vistes dels objectes.

5.1 Obtenció del mapa de disparitat.

En aquest apartat es mostrarà com crear el mapa de disparitat per diferents vies: la d'aparellament estèreo de punts semblants entre imatge i les derivades de la coneixença per altres camins de l'estructura tridimensional de l'escena.

5.1.1 Aparellament estèreo.

Per obtenir el mapa de disparitat entre dues imatges és necessari saber quin píxel de la imatge 2 correspon a un donat de la imatge 1, i per tant, cal fer una cerca. A priori no es pot saber a quina zona de la segona imatge es trobarà el píxel corresponent a no ser que es tinguin condicions geomètriques especials en que es garanteix que la cerca es farà en un segment determinat. Aquestes condicions s'obtenen amb una disposició epipolar de les càmeres o amb un procés de rectificació que garanteixi la coplanaritat de les imatges reprojectades (veure figura 5.2).

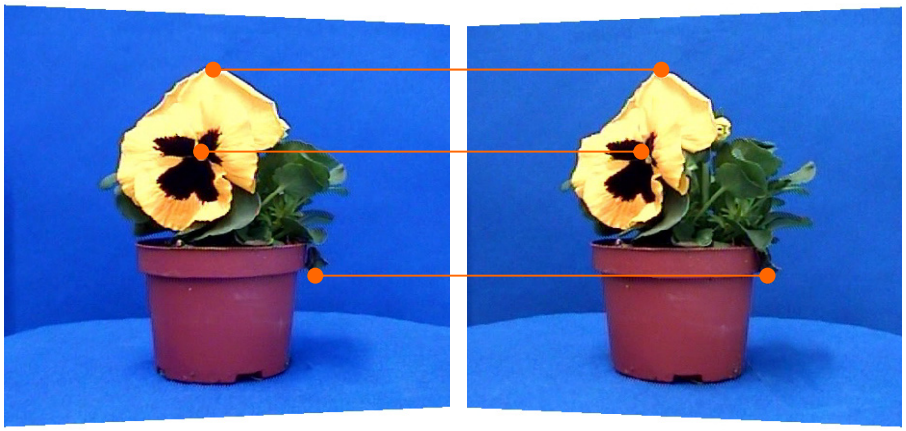


Figura 5.2 Parell d'imatges rectificades per obtenir unes condicions d'aparellament de punts com les de la geometria epipolar.

En qualsevol cas, cal fer, per cada parell d'imatges, una cerca per obtenir el mapa de disparitat. Hi ha una gran varietat d'algoritmes per a correspondència estèreo, molts d'ells es poden consultar en el treball de Selizky i Scharstein [Scharstein 02], i en l'estat de l'art d'aquesta mateixa tesi.

A l'hora de computar com de semblants són dos grups de píxels, un dels mètodes més emprats és l'anomenat SSD (de l'anglès *sum of squared diferences*), que computa la suma de diferències al quadrat entre un grup de píxels, centrat en el píxel de la imatge original, i un grup de píxels que va desplaçant-se per l'altra imatge. La disparitat on aquesta suma de diferències es fa mínima es pren com la millor correlació possible.

Així doncs, l'algoritme de cerca de corresponents amb SSD es pot caracteritzar així:

```
per tots els píxels de la imatge primera
{
        per una finestra de mida definida
        {
                per un rang de disparitats donat
                {
                        seleccionar aquella disparitat que dóna un menor
                        error segons la SSD en la finestra centrada en
                        el píxel de la segona imatge dins el rang
                }
        }
}
```

El seu cost computacional és doncs de l'ordre de $n \cdot d \cdot w$ on n és el nombre de píxels de la imatge, d el rang de disparitats a cercar i w la mida de la finestra de cerca. Els mapa computat a les figura 5.1 s'ha realitzat amb aquest algoritme.

El principal problema dels algoritmes com el SSD és que sols es pot buscar efectivament el píxel corresponent d'un donat en les regions on existeix gradient a la imatge, per les regions homogènies de la imatge caldrà estendre o interpolar les disparitats obtingudes en les zones amb gradient (veure figura 5.3).

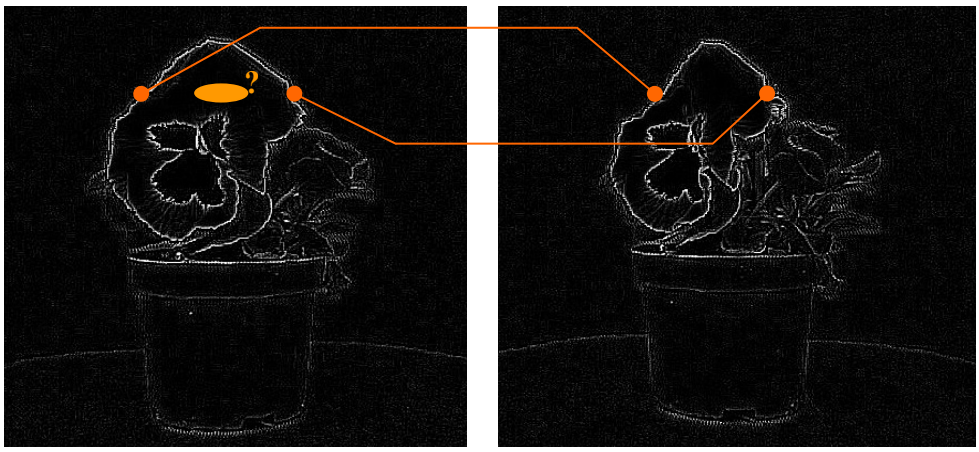


Figura 5.3 Els algoritmes de cerca de correspondència com el SSD només troben aparellament en els punts on hi ha gradient.

Tot això fa que altres algoritmes més costosos com els de programació dinàmica, ajust de superfícies bidimensionals genèriques damunt l'espai a reconstruir, tècniques amb equacions diferencials, minimització d'errors , optimització, SMP (*single matching phase*) o BM (*bidirectional matching*) puguin obtenir millors resultats, amb mapes de disparitat molt més densos que l'obtingut per SSD. Això si, a un cost computacional encara més alt que pot fer incompatibles molts d'ells amb qualsevol aplicació de vídeo sota demanda o realitat augmentada interactiva. Al capítol d'estat de l'art de la tesi es referencien alguns d'aquests mètodes, i s'han fet proves amb algun d'ells, concretament amb programació dinàmica [Martín 03] però cap dels mètodes d'aparellament estèreo

no assoleix tanta qualitat com la que es pot obtenir a partir d'altres tècniques de reconstrucció tridimensional de l'objecte.

Problema plantejat:

Aquest compromís entre qualitat (avaluada en funció de la densitat del mapa i del seu error) i cost computacional, del còmput del mapa de disparitat, va fer imperatiu cercar algun mètode que permetés tenir una dada de disparitat correcta en el mínim temps possible. Amb condicionaments com el de l'algoritme SSD, el mètode de síntesi de vistes vist al capítol 4 que requereix l'ús del mapa de disparitat extensivament no podrà generar imatges de qualitat de manera interactiva.

Condicions de la solució a cercar:

Donat que es vol estudiar l'ús de les tècniques de síntesi de vistes en aplicacions on s'hauria pogut fer certa computació prèviament (processos diferits en el temps), s'ha ampliat el ventall de solucions avaluades a d'altres que poden arrossegar certa informació de l'objecte abans de fer la síntesi i:

1) La millor solució trobada ha estat la de tenir precalculada informació suficient de la correspondència entre els píxels de les vistes com per poder generar el mapa de disparitat eficaçment en el procés de síntesi de les noves imatges.
2) Òbviament, aquesta solució només serà aplicable en el cas d'haver adquirit prèviament la informació fotomètrica que s'usarà i coneixent-ne els punts de vista; si les imatges s'estan adquirint en temps real, caldrà computar contínuament els mapes de disparitat.
3) Donat que es poden demanar noves vistes des d'ubicacions arbitràries, no és possible precalcular directament el mapa de disparitat, ja que en el procés de rectificació els seus valors perdrien sentit.

A continuació es mostra quin camí s'ha triat per cercar una solució, amb quin criteri s'ha fet la selecció i es comentarà com es pot enregistrar la informació necessària per generar dinàmicament el mapa de disparitat.

5.1.2 Informació tridimensional, disparitat i correspondència.

Com s'ha dit, existeix una relació entre els valors de disparitat dels píxels i la profunditat del punt referenciat. Donada la posició tridimensional d'un punt, es projecta aquesta sobre les dues imatges i es tenen les coordenades dels píxels projectats. Si es posen en una estructura de dades les parelles de píxels originades per un mateix punt, s'obté un mapa de correspondència. Si el que es fa es calcular les distancies en el pla de reprojecció de les dues imatges el que s'obté és el mapa de disparitat. Així doncs es tenen tres maneres equivalents d'expressar la idea d'estructura tridimensional:

- Explícitament (1), on cada element representa una coordenada X,Y,Z d'un punt de l'objecte a l'espai, en aquest cas la distància s'obté directament.

- Implícitament amb el mapa de disparitat (2), que per cada píxel guarda informació de la distància al seu corresponent a l'altra imatge. Amb aquestes dues posicions i informació de distància entre les càmeres es pot triangular la posició dels punts en 3D.
- Implícitament guardant les correspondències entre punts (3). Dels píxels de la imatge original es guarden les coordenades del seu corresponent a l'altra imatge. A partir d'aquests enllaços es pot calcular fàcilment el mapa de disparitat per seguir com en l'apartat anterior.

La figura 5.4 mostra la relació entre les diferents opcions exposades per tenir la informació tridimensional de les imatges que s'utilitzarà per generar els mapes de disparitat.

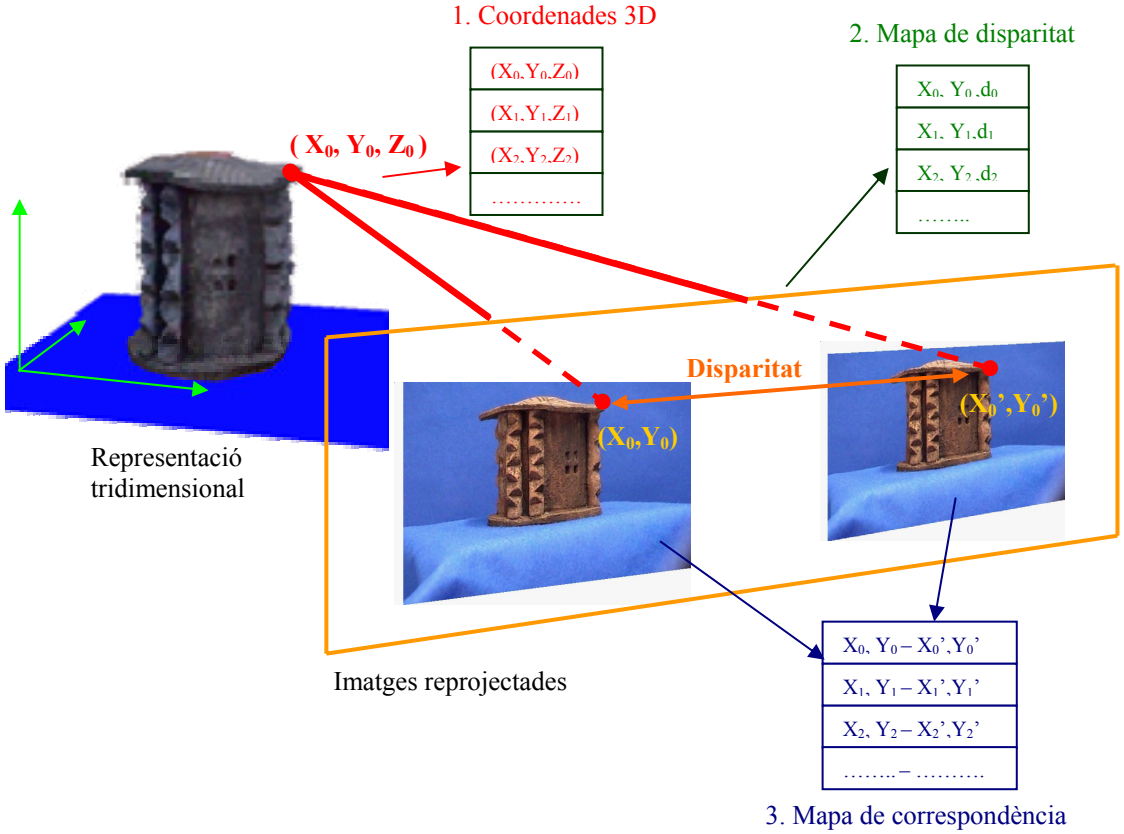


Figura 5.4 Tres maneres diferents de representar la informació tridimensional: llista de punts explícita en 3D, mapa de disparitat i mapa de correspondència, i la manera com es representen i relacionen entre ells a l'espai.

Així doncs si es suposa que es té o s'ha reconstruït el model tridimensional de la escena observada, es plantegen tres opcions per emmagatzemar aquesta informació: gravar l'estructura tridimensional juntament amb les imatges, gravar el mapa de disparitat juntament amb les imatges i la tercera, gravar el mapa de correspondència juntament amb les imatges. A continuació s'avaluaran els avantatges i inconvenients de cada una d'aquestes possibilitats per a implementar-se en un sistema interactiu de realitat virtual o telepresència que utilitzi tècniques de síntesi de vistes:

1) Gravar imatges i informació tridimensional

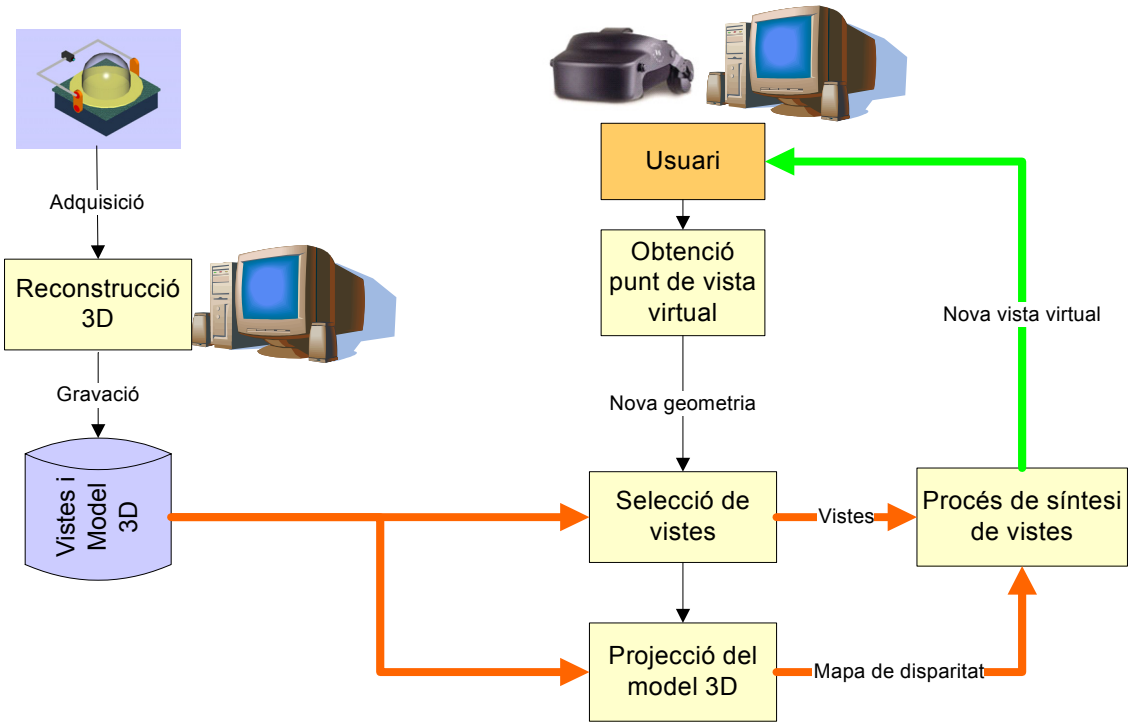


Figura 5.5 Obtenció del mapa de disparitat per a la síntesi de vistes utilitzant un model tridimensional dels objectes.

En aquesta opció s'emmagatzemarà informació sobre les vistes de l'objecte i el model tridimensional del mateix. D'aquesta manera quan el procés interactiu determina la nova geometria es seleccionen les vistes més properes i s'obté el mapa de disparitat projectant sobre els punts de vista rectificats la informació tridimensional, en el nou espai la disparitat serà la resta de les components horitzontals de les projeccions. A partir de les vistes i el mapa de disparitat, es sintetitza la vista tal com s'ha mostrat en el capítol quart

Avantatges:

El fet de tenir la informació tridimensional fa que el càlcul del mapa de disparitat sigui senzill, consistint en projectar els punts a les dues càmeres en el nou pla, i restar les components horitzontals.

Inconvenients:

Com que per cada parell de vistes es té el model sencer, existeix un problema d'ordenació espacial a l'hora de projectar. Per resoldre'l es pot utilitzar la tarja gràfica i llavors recuperar la informació, però el procés es fa més lent. D'altra banda, si ja es té el model tridimensional explícit, es pot plantejar algun mètode alternatiu a la síntesi de vistes, tal com es veurà al final d'aquest capítol.

2) Gravar imatges amb el mapa de disparitat

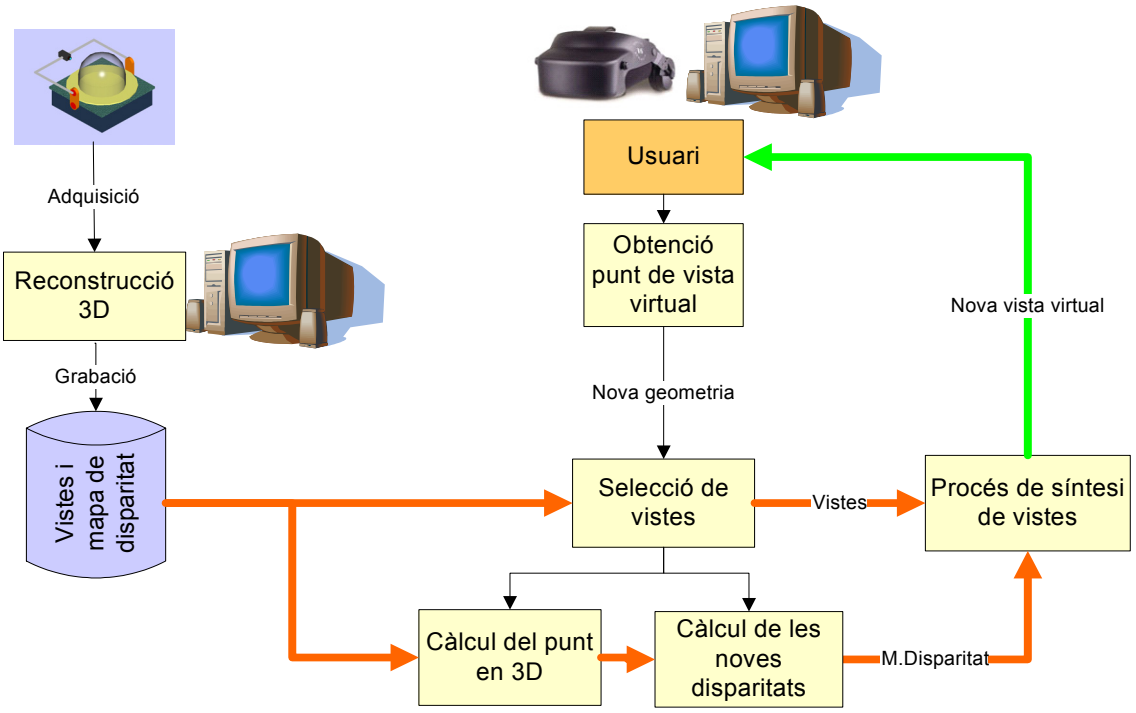


Figura 5.6 Obtenció del mapa de disparitat per a la síntesi de vistes utilitzant mapes de disparitat calculats prèviament.

En aquesta opció es té a disc informació sobre les vistes de l'objecte i els mapes de disparitat entre parelles de vistes. Tot i que sembli òptim, el fet de que al triar el punt de vista virtual calgui reprojectar les imatges, implica que per cada parell de punts cal retrobar la posició tridimensional i projectar-la en la nova geometria.

Avantatges:

Com que es té el mapa de disparitat calculat, en les poques configuracions en que no variï la geometria de l'escena, es podrà passar ràpidament al procés de síntesi i representació.

Inconvenients:

El fet de que en moltes ocasions calgui recalcular el mapa de disparitat, implicant un nou procés de trobar l'estructura tridimensional, fa inútil el fet d'haver-lo trobat prèviament. Aquest mètode no presentarà cap avantatge respecte al primer o al tercer en la majoria dels casos, i per tant es descarta.

3) Gravar imatges i mapa de correspondència

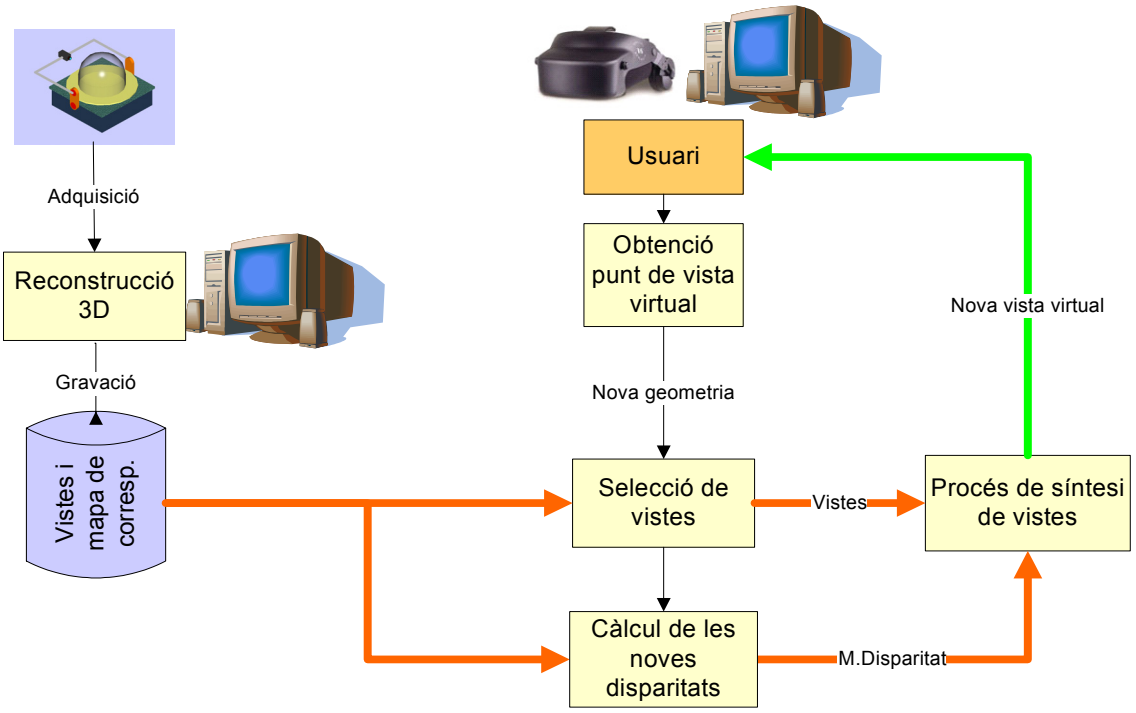


Figura 5.7 Obtenció del mapa de disparitat a partir dels mapes de correspondència.

En aquesta cas el que es guarda al disc és informació sobre les vistes de l'objecte i els mapes de correspondència entre píxels de parelles de vistes. Quan s'ha determinat la nova geometria, cada píxel original i el seu corresponent a l'altre imatge són portats mitjançant el producte per les matrius d'homografia (veure capítol 4) al nou pla de projecció. Un cop allà s'obté el mapa de disparitat restant les components horitzontals i ja es pot fer la síntesi de vistes.

Avantatges:

Amb el mapa de correspondència, trobar la disparitat entre els píxels, tant en l'espai original com en el rectificat és tan senzill com seguir un enllaç informàtic. Això fa que aquesta sigui la opció més eficient de les tres plantejades ja que en cap cas caldrà explicitar el 3D dels punts de l'escena.

Conclusió de l'estudi del pas de la informació tridimensional a disparitat.

S'ha cregut doncs que gravar la informació de les vistes juntament amb els mapes de correspondència serà la millor opció per poder aplicar ràpidament el procés de síntesi de vistes. Malgrat tot, en els tres casos es té la necessitat de tenir una bona reconstrucció tridimensional de l'escena. La qualitat de les imatges obtingudes, sigui pel mètode que sigui, serà directament proporcional a la bondat de les dades tridimensionals. El següent apartat mostra la problemàtica de la reconstrucció tridimensional i els camins seguits per obtenir-la en els experiments realitzats.

5.2 Reconstrucció tridimensional

Existeixen diversos mètodes per l'obtenció de la informació tridimensional dels objectes o escenes; alguns utilitzen únicament la informació (imatges) donada per les càmeres, a vegades calibrades, a vegades no. Si les càmeres no estan calibrades sorgeix el problema addicional de determinar els paràmetres de la càmera; per això s'empren algoritmes d'autocalibració, de calibració per conjunts de punts aparellats (per trobar la matriu fonamental del sistema [Faugeras 93]) i altres. Amb les dades fotomètriques dels píxels es poden aplicar algoritmes d'aparellament estèreo, que van resolent individualment posicions de punts a l'espai, de *view consistency*, *voxel coloring* o *space carving*.

Tots aquests mètodes presenten dificultats d'implementació i restriccions en la naturalesa de l'objecte com la necessitat d'ordenació esquerra-dreta dels punts, la no tolerància a oclusions o el comportament del material a la llum (en general, els materials dels objectes tractats han de ser no transparents i no especulars).

Per tot això s'ha cregut convenient disposar d'un mètode alternatiu d'obtenció de l'estructura tridimensional. D'aquesta manera es podran comparar els resultats dels altres mètodes emprats i avaluar-ne l'error. Aprofitant que s'ha construït un sistema robotitzat per l'adquisició de les imatges, s'ha decidit d'incorporar-hi un pla làser que facilita la identificació dels píxels amb la càmera i l'obtenció de la seva ubicació en tres dimensions per triangulació. A continuació es mostra de manera didàctica la incorporació del pla làser al sistema, el resultat obtingut amb el làser, l'avaluació dels mètodes de *voxel coloring* i de *space carving*, que finalment s'ha emprat per obtenir la informació tridimensional de l'objecte i al que s'ha dedicat el capítol sisè.

5.2.1 Reconstrucció tridimensional emprant un pla làser i una càmera.

Donat el dispositiu d'adquisició descrit al capítol tercer, s'hi ha instal·lat un projector d'un pla làser que formarà un angle constant de trenta graus amb l'eix òptic. La figura 5.8 mostra tres fotografies obtingudes en el procés d'adquisició on s'hi aprecia la projecció del pla làser en vermell.

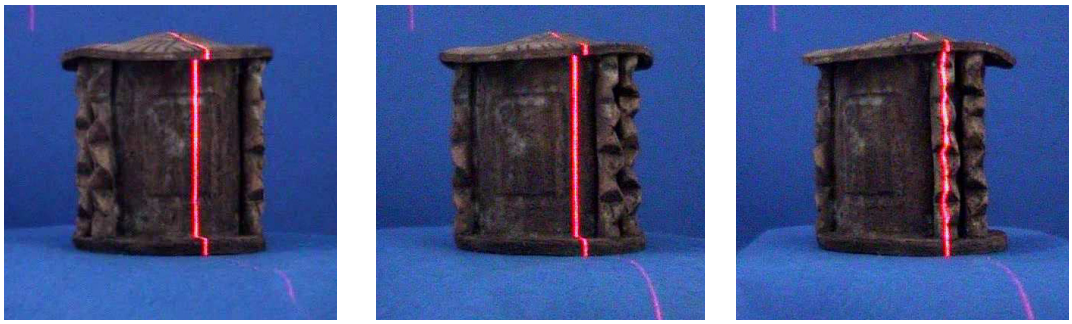
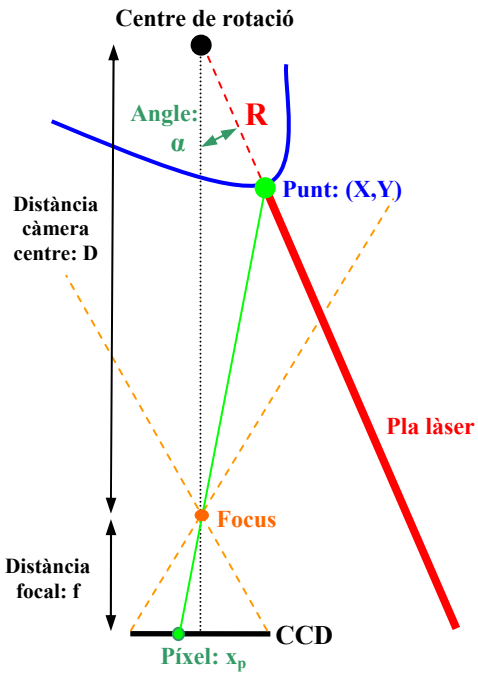


Figura 5.8 Imatges del procés de reconstrucció damunt una urna funerària romana, gentilesa del departament d'arqueologia i història antiga de la Universitat de Barcelona.

Un cop obtingudes les imatges amb el làser projectat, un senzill procés de triangulació, descrit a la figura 5.9 permet obtenir la posició tridimensional del punt a l'espai. La repetició d'aquest procés per tots els punts de la línia i per totes les línies de la seqüència d'imatges permet obtenir el model de l'objecte.



Tenint

D: distància de la càmera amb el centre de rotació
f: distància focal
α: angle eix òptic - làser.

$$R = \frac{f \cdot x_p - D \cdot x_p}{f \cdot \sin(\alpha) - x_p \cdot \cos(\alpha)}$$

I es troba:

$X = R \cdot \sin(\alpha)$
$Y = R \cdot \cos(\alpha)$

Figura 5.9 Procés de triangulació per obtenir les dimensions de l'objecte a partir de la projecció del làser, la figura mostra com a partir de la coordenada x_p del píxel es troben les coordenades X i Y del punt a l'espai. La coordenada Z es troba a partir de la y del píxel amb l'equació del model *pin-hole*.

Resultats.

Un cop fet el recorregut de tots els píxels, de totes les vistes de la seqüència s'obté el resultat en forma de matriu de punts tridimensionals. La matriu tindrà tantes files com vistes tingudes de l'objecte (és a dir com passos en la volta horitzontal del posicionador) i tantes columnes com píxels en la imatge. La dada guardada en cada posició de la matriu serà les coordenades [x, y, z] del punt a l'espai (veure la taula 5.4). Es pot ja, a partir d'aquestes dades, generar la llista de punts corresponents entre vistes donades.

Evidentment, també es pot aprofitar la capacitat dels ordinadors personals de dibuixar imatges tridimensionals: la figura 5.10 mostra la representació amb una llibreria gràfica dels punts tridimensionals damunt una càmera virtual.

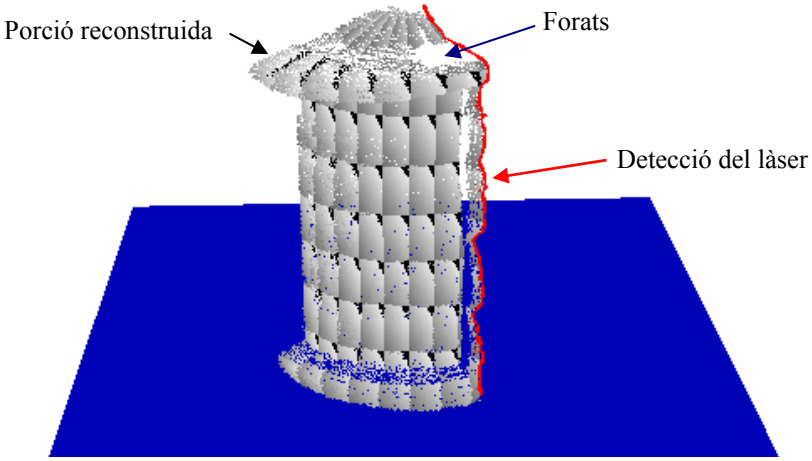


Figura 5.10 Instant del procés de reconstrucció de l'estructura tridimensional mitjançant la detecció del làser i la triangulació dels punts

Aquest és un exemple de l'anomenat procés de *render* on, aprofitant els recursos de la targeta gràfica, GPU i memòria, es pot obtenir qualsevol vista de l'objecte. Es planteja doncs una alternativa als dos processos d'obtenció de vistes mostrats fins el moment: el d'agafar un model tridimensional, donar-li la textura de l'objecte en qüestió i dibuixar-lo. La figura 5.11 mostra el resultat d'aquest procés. Es combina la informació tridimensional que ofereix la triangulació amb el làser amb la de color que s'obté amb una càmera normalment. Els forats com els representats a la figura 5.10 han estat interpolats linealment. En el capítol de resultats es comparará aquest mètode amb el de selecció de vistes i el de síntesi de vistes.

Figura 5.11 Representació tridimensional del model de l'objecte reconstruït i interpolat.

Precisió, fonts d'error.

És important plantejar-se com de precisa és la reconstrucció que s'ha mostrat amb la triangulació de la càmera i el làser. Per una banda cal considerar quina seria la resolució màxima que es pot obtenir, per altra quines fonts d'error afectaran al sistema i finalment, que s'ha fet per intentar minimitzar aquests errors.

Com s'ha vist en el capítol tercer (veure taula 3.1) el posicionador donarà una resolució horitzontal de 0,31 mm per un objecte de 6 cm de radi com el reconstruït i una resolució vertical de 0,23 mm. Com que l'adquisició es fa sempre en un procés seqüencial i el moviment és sempre relatiu a la posició anterior, l'error de precisió absolut podrà ésser deixat a banda. Queda ara veure quina resolució tindrà la càmera en la detecció del làser i els errors relatius a la geometria.

El procés de càlcul es basa en la triangulació de la posició dels punts de l'objecte il·luminats pel làser i capturats per la càmera. En ell es suposa que:

- L'eix òptic de la càmera i el pla làser es tallen en un punt.
- El punt de tall es troba damunt de l'eix de rotació del posicionador.
- L'eix de rotació del posicionador està contingut pel pla làser.
- L'angle format pel pla làser amb l'eix òptic és conegut i no varia amb el moviment del posicionador.

- El plat del posicionador és pla i perpendicular a l'eix de rotació.
- El centre de masses de l'objecte està el més proper possible a l'eix de rotació del posicionador i s'assumeix que les concavitats de l'objecte provocaran forats en la reconstrucció que s'hauran d'interpolar.

Qualsevol error en alguna de les cinc primeres condicions es traslladarà a la reconstrucció tridimensional de l'objecte i en minvarà la precisió. A continuació es llistaran les mesures preses per intentar minimitzar aquests errors:

- El sistema es calibra per determinar l'eix de rotació del posicionador, aquesta mesura es fa òpticament amb la mateixa càmera i amb uns punts far que es van seguint en el procés.
- Donada la visibilitat del làser es fa servir la càmera per determinar l'angle incident i la verticalitat del pla projectant-lo damunt diferents superfícies a distàncies conegudes.
- Es col·loquen objectes de dimensions molt ben conegudes damunt del plat per comprovar que gira uniformement i que és pla. Si no ho fos, es trobaria que la cota màxima de l'objecte projectat a la càmera varia amb la rotació.

Donat que durant la calibració totes les mesures s'han efectuat amb la càmera, es pot estimar que la precisió serà aproximadament de la mida d'un píxel, és a dir, per la càmera emprada de 720 píxels horitzontals i amb el camp d'enfocament emprat en la calibració, es pot estimar en 0.12 mm. Així doncs la posició d'un punt de l'objecte a l'espai serà determinat amb una resolució de 0.3 mm ± 0.12 mm.

5.2.2 Ús dels mètodes de selecció de vòxels.

Alternativament als mètodes on es triangula explícitament la posició d'un punt a l'espai a partir d'un píxel vist per dues càmeres o d'un punt vist per una càmera i il·luminat per un làser, existeixen mètodes amb els que es genera un volum de *voxels* igual al volum de l'escena a priori i es van eliminant aquells que estan buits. És important referir-se a aquests mètodes ja que la majoria d'objectes no podran ser il·luminats amb un làser en un robot posicionador o la reflexió del làser no podrà ser detectada. En aquesta tesi s'ha decidit usar en paral·lel els dos mètodes: reconstrucció amb làser i per selecció de vòxels, per poder comparar els errors comesos o millorar el model fusionant les estructures obtingudes.

Amb les tècniques de selecció de vòxels s'obté un conjunt (s'intentarà que el mínim) de vòxels que conté els objectes de l'escena. Aquest volum es pot representar sencer o es pot cercar només la carcassa envoltant dels objectes o en anglès *convex hull*. Per seleccionar els *voxels* de l'escena que quedaran representats existeixen dues tècniques bàsiques: l'acolorit de voxels amb comprovació de coherència en les vistes i l'esculpit de l'espai a partir de la projecció dels objectes a les vistes. Així doncs, donat el conjunt de vistes obtingut i una discretització del volum de treball del posicionador, es podrà obtenir l'estructura tridimensional de l'objecte.

5.2.2.1 *Voxel coloring.*

El mètode d'acolorit de vòxels o en anglès, *voxel coloring*, consisteix en, per cada vòxel del volum de treball, cercar la projecció a totes les vistes i si, la llista de colors trobats és coherent deixar-lo i sinó eliminar-lo. La computació d'aquesta "coherència" ofereix diverses possibilitats però, comunament, s'analitzen estadísticament els valors de la llista de colors trobats. Si existeix una distribució normal entorn a un valor de color, es suposa que aquest és correcte i s'assigna al vòxel en qüestió. Si la dispersió dels valors és massa gran, el vòxel s'elimina de l'estructura de dades (veure figura 5.12).

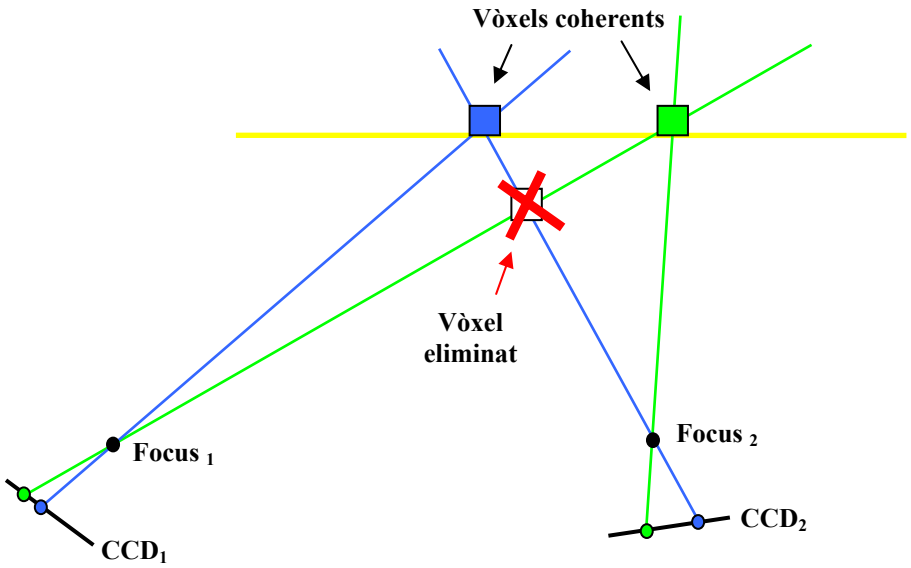


Figura 5.12 Representació esquemàtica del mètode de *voxel coloring* per dues càmeres observant una escena.

El cost algorísmic típic d'aquest mètode és (suposant que es tenen N^3 vòxels i M vistes d'un objecte): $N^3 \cdot M$ projeccions + N^3 processats d'una llista de M elements. A priori és un cost molt gran, tot i que el fet de que el procés es pugui fer diferidament (en anglès *batch*) trauria importància a aquest cost.

5.2.2.2 *Space carving.*

El mètode d'esculpit de l'espai o esculpit del volum de vòxels consisteix en, per totes les vistes de l'objecte, projectar-hi els vòxels i, aquells que cauen en la regió de la imatge amb píxels de l'objecte (*foreground*) mantenir-los, i la resta (*background*) eliminar-los. La figura 5.13 mostra el funcionament del procés per una vista donada. Caldria repetir (en principi) aquest procés per totes les vistes disponibles de l'objecte. L'obtenció de les regions segmentades de la imatge es pot fer paral·lelament a l'adquisició ja que en tots els casos tractats s'ha posat un fons fàcilment distingible per selecció de color (en anglès *chroma key*).

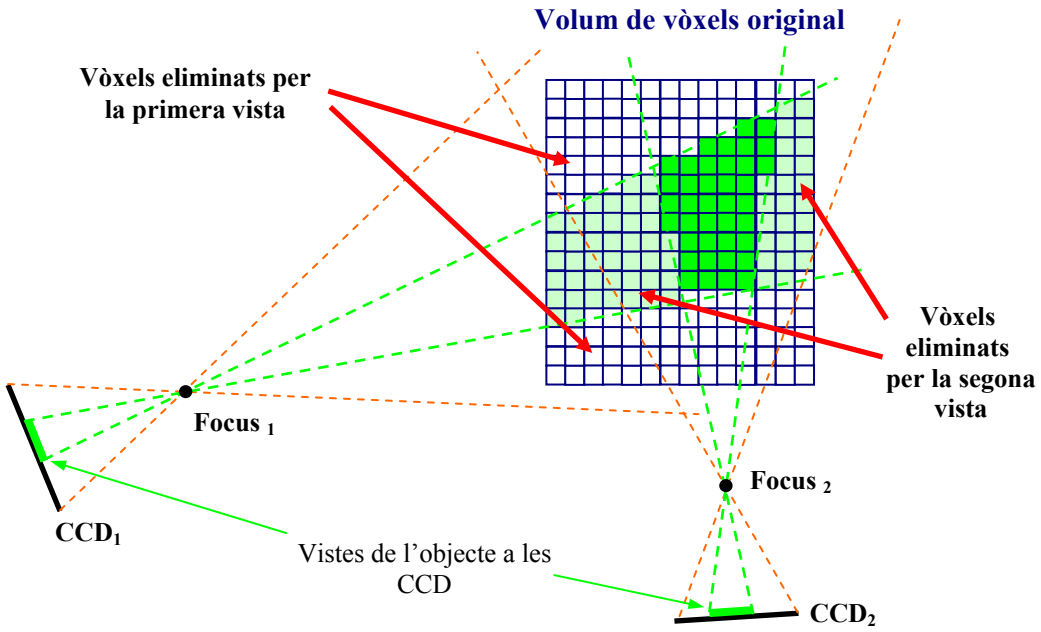


Figura 5.13 Representació esquemàtica del mètode de *space carving* amb dues càmeres observant un volum de vòxels.

El cost algorísmic màxim d'aquest mètode és (suposant que es tenen N^3 vòxels i M vistes d'un objecte): $M \cdot N^3$ projeccions + M operacions lògiques damunt dels píxels de la imatge projectada. En qualsevol cas, cal considerar que:

- Per la primera vista hi ha N^3 *voxels* a projectar, dels que se n'eliminaran molts, que ja no caldrà projectar a les posteriors i així successivament fins a la vista M-èssima.
- L'ordre en que es projecten les vistes farà reduir encara més el cost total del mètode i potser es podrà trobar un conjunt suficient de vistes menor que M a projectar.
- La projecció dels *voxels* (això també és aplicable al cas de *voxel coloring*) es podrà fer aprofitant els recursos del *hardware* (maquinari) de la VGA.
- La selecció dels *voxels* a eliminar es podrà fer amb una operació lògica entre imatges (imatge projectada *and* vista original binaritzada) que també pot ser realitzada ràpidament per la VGA.

Per tot això s'ha considerat apropiat aplicar el mètode de *space carving* per l'obtenció del model tridimensional de l'escena. Donat que sobre el mètode genèric s'ha decidit efectuar-hi diverses millores, es dedica un capítol sencer, el sisè, a tractar com fer de manera òptima aquest esculpit de l'espai com seleccionar les vistes necessàries per a fer-lo i quin suport pot donar el maquinari d'un PC per fer-ho amb el mínim cost computacional.

6. *Space Carving*. Selecció de punts de vista

En aquest capítol es tractarà específicament el tema de la selecció de vistes. En el capítol quart s'ha descrit un procés que, a partir d'un parell de vistes i el mapa de disparitat pot sintetitzar les vistes intermèdies. En el capítol cinquè s'ha vist que per obtenir el mapa de disparitat cal disposar d'informació tridimensional de l'escena i que el mètode de *space carving* pot extreure-la a base d'eliminar els vòxels no projectats a les diferents vistes. En ambdós casos es pot fer la pregunta: quines vistes s'han d'utilitzar? totes? algunes? amb quin criteri? En aquest capítol es cercarà resposta a aquestes preguntes de manera que es pugui acabar de definir de manera òptima el procés de *space carving* i el mètode d'obtenció de vistes per selecció i síntesi. També s'aprofitarà per introduir certes millores en el mètode de *space carving* tant a nivell temporal ja que, en general, és un mètode molt costós en temps de processador, com en la qualitat del model resultant.

6.1 Tractament de les imatges per al *carving*.

En primer lloc es veurà com preparar el conjunt d'imatges original per al procés de *space carving*, com s'ha vist en el capítol tercer, inicialment es té una llista d'imatges ordenades per el punt de vista des de el que han estat preses. El coneixement del punt de vista serà utilitzat per al procés de projecció, la informació fotomètrica de l'escena es farà servir per distingir el *foreground* (part interessant) i el *background* (fons) amb un procés de segmentació. Un cop s'ha segmentat la imatge s'hi apliquen operadors morfològics per filtrar petits errors i se n'extreu la característica àrea de l'objecte segmentat, que serà utilitzada més endavant en aquest capítol. Tots aquests són processos clàssics de visió per ordinador i només s'exemplificarà breument el seu ús

Segmentació. *Chroma key*.

A l'hora de plantejar-se la plataforma d'adquisició d'imatges es va decidir fer tot el possible per facilitar el procés de segmentació de l'objecte d'interès respecte al fons. En un entorn no controlat, caldria separar l'objecte del fons o altres objectes amb un procés de etiquetatge i identificació dels elements de la escena, detecció de les parts mòbils respecte al fons o altres. En el cas dels objectes usats en la fase d'experimentació d'aquesta tesi s'ha col·locat sempre un fons de color molt diferent al presentat per

l'objecte d'interès. D'aquesta manera, el procés de segmentació ha pogut ésser simplement una comparació entre el color conegut de fons i el color del píxel: si el color del píxel és molt semblant al fons s'elimina i en cas contrari es deixa. Aquest és el procés conegut com a *chroma key* o selecció cromàtica. Queda determinar la mesura de similitud entre dos píxels; la representació obtinguda en tres bandes verd, blau i vermell convida a utilitzar una mesura com la distància euclidiana però està demostrat que es més òptim treballar en l'espai transformat HLS (de *Hue Light Saturation*, traduïble com tonalitat, lluminositat i saturació de color). En conseqüència s'ha realitzat aquesta transformació i segmentat l'objecte per distància entre components H dels píxels respecte al fons (veure figura 6.1).

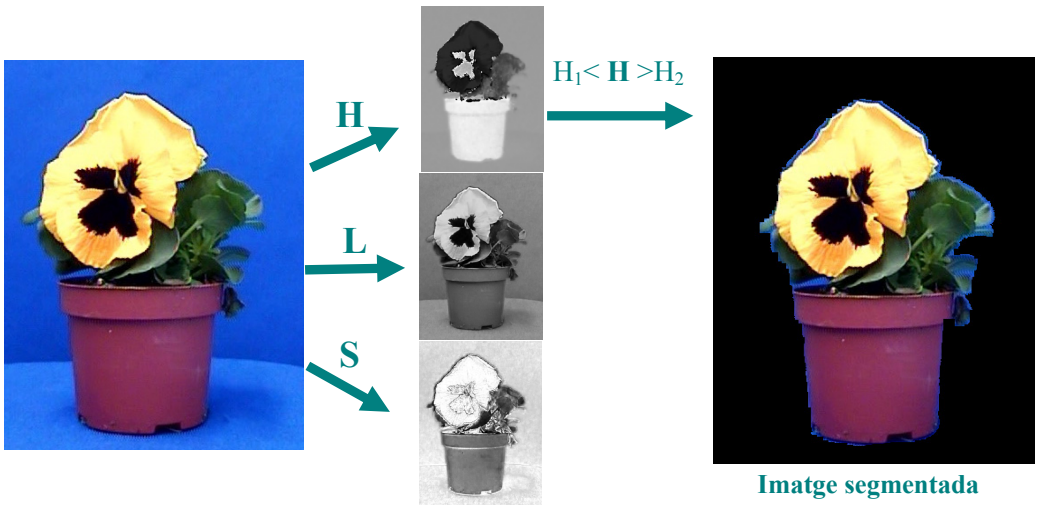


Figura 6.1 Procés de segmentació: conversió al model HLS i eliminació del fons en la banda H (tonalitat de color).

Processat. Característica àrea.

Un cop segmentat l'objecte s'han aplicat operadors morfològics, concretament un tancament per eliminar punts erronis en el *foreground* degut a soroll o reflexes. L'únic processat posterior que s'ha fet ha estat, per totes les imatges preses de l'objecte amb el posicionador, extreure la característica àrea, com es representa en la figura 6.2 i que s'usarà posteriorment.

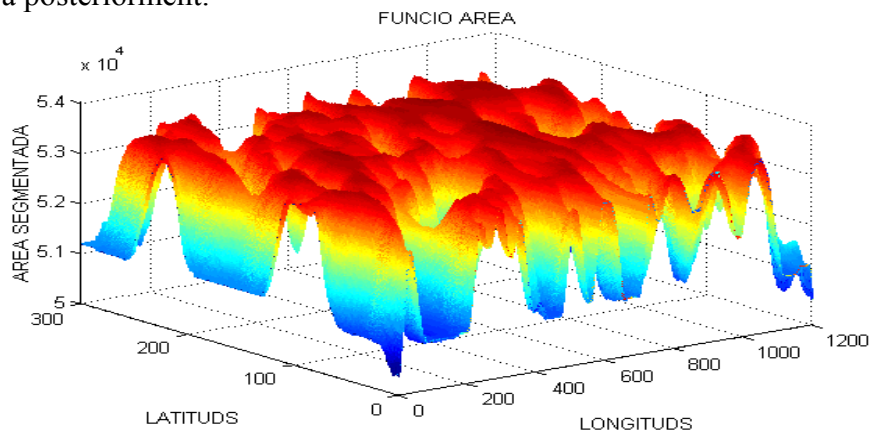


Figura 6.2 Vista de la representació tridimensional de la funció àrea de la imatge segmentada, respecte a l'índex de les vistes horitzontals i verticals.

6.2 Millores en la codificació i projecció dels voxels.

La representació del volum de vòxels en la memòria del processador es pot plantejar eficientment de dues formes: com una matriu tridimensional de vòxels M[i][j][k] o com una estructura en arbre on de cada node surten vuit fills representant la divisió de l'espai contingut en vuit cubs d'igual mida, de cada fill igualment surten vuit més fins arribar a la divisió mínima de l'espai (model de *octree*). La primera opció és més senzilla a nivell algorísmic i la segona més eficient en l'ús de memòria. Si es té una representació matricial cada element és un vòxel a projectar i del qual caldrà decidir si pertany o no a l'objecte. Amb la representació en arbre es poden eliminar centenars de cel·les filles quan es determina que un node no ha de ser representat, però la projecció de blocs grans és més difícil de computar. Per determinar les dimensions de l'arbre o la mida de la matriu cal establir la mida mínima del vòxel representat i dividir l'espai de treball per aquesta mida i aquí sorgeixen diversos problemes ja que l'espai definit pel posicionador (o per la rotació d'una càmera entorn un objecte qualsevol) és cilíndric, la projecció a la càmera és cònica i la manera més senzilla de representar els vòxels és un espai ortonormal (veure figura 6.3). Desgraciadament no s'ha trobat cap solució millor que la trivial que és representar els vòxels en un espai ortogonal i assumir els errors que això implica.

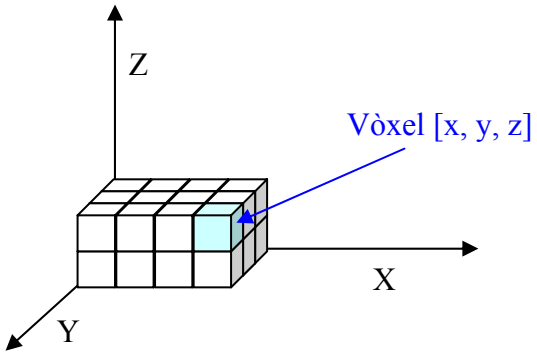


Figura 6.3 Definició estàndard del volum de vòxels en un espai ortogonal.

6.2.1 Us dels recursos gràfics per al *carving*.

El procés de *space carving* es pot plantejar de dues formes: projectant les imatges damunt els vòxels o projectant els vòxels damunt les imatges.

El primer cas implica, per tots els punts del fons de la imatge, traçar rectes a l'espai i calcular la distància entre els vòxels i les rectes. Per un vòxel donat, si la distància és menor a la meitat de la mida del vòxel aquest és eliminat. És un procés costós que ha de realitzar principalment el processador i amb un cost del ordre del nombre de píxels pel nombre de vòxels.

Definint les variables "mida_V" com el costat del cub de vòxels i "mida_Im" com el nombre de píxels de la imatge, el cost de l'algoritme és:

$$\text{Cost (traçat rectes)} = k_0 + k_1 \cdot \text{mida\_Im} \cdot (\text{ mida\_V })^3 \qquad \text{(Eq. 6.1)}$$

En el segon cas cal projectar els vòxels damunt les imatges i veure si el punt destí pertany a l'objecte o al fons. Aquest mètode, presenta diverses avantatges respecte al primer, ja que es pot aprofitar el maquinari de la targeta gràfica existent als ordinadors personals per fer totes aquestes operacions:

- Acceleració per hardware del procés de projecció.
- Paral·lelització del procés de projecció dels vòxels.
- Realització de la operació de tria del vòxel amb una operació lògica entre imatges.

Per implementar aquestes millores, és necessari aplicar el següent algoritme:

```
#definir mida_V // amplada del cub de vòxels
#definir mida_Im // mida de la imatge en píxels

/* Construir vòxels */

per ( i = 0; i<mida_V ;i ++ )
per ( j = 0; j<mida_V ;j ++ )
per ( k = 0; k<mida_V ;k ++ )
        CrearVoxelColor( i, j, k);              //  crea un vòxel a la posició i, j, k
                                                //  amb color R,G,B = i, j, k

/* Carregar imatge vista */
        im_bin = CarregaImatge( );              // carrega la imatge segmentada amb
                                                // background=1 i foreground=0
/* Procés de space carving */
per (x = 0; x<mida_V; x++)
{
        im_rend = ProjectaVoxels( );            // fa la projecció de l'escena de vòxels
        im_result = im_bin & im_rend;           // operació and lògica

        per ( y = 0; y<mida_Im ; y++)
                si ( im_result [y] != 0 )       // si el píxel és del fons
                        EliminaVoxelColor (im_result [y] )

}
```

On el punt clau és l'assignació dels colors als vòxels; el fet de que cada vòxel tingui un color únic que l'identifica permet de manera senzilla projectar la seva estructura com un gràfic ordinari, fer la màscara amb la imatge binaritzada, i identificar directament els vòxels que han de ser eliminats. El cost resultant d'aquesta implementació, amb l'avantatge de que serà executat en gran part pel coprocessador gràfic és:

$$\text{Cost(projecció)} = k_0 + k_1 \cdot (\text{mida\_V}^2 + \text{mida\_Im}) \cdot \text{mida\_V} \qquad \text{(Eq. 6.2)}$$

On per les equacions 6.1 i 6.2 els paràmetres k_i representen:

k_0: cost d'inicialització d'estructures.
k_1: cost de la decisió de selecció de vòxel.
mida_V^2: cost de la projecció d'una estructura al coprocessador gràfic.

Experimentalment es troba que el segon mètode, que usa el coprocessador gràfic, amb maquinari especialitzat per a les projeccions, ofereix millor rendiment que el primer (veure gràfic a la figura 6.4) . La figura 6.5 il·lustra el funcionament d'aquest

segon mètode, amb la codificació de vòxels amb colors i operació lògica amb la imatge binaritzada. A continuació es planteja un mètode alternatiu sobre el procés de representació i projecció de vòxels per al mètode de *space carving*.

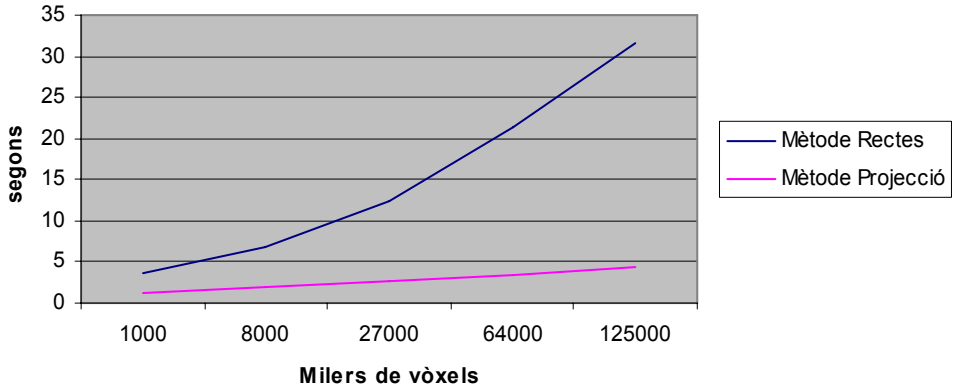


Figura 6.4 Gràfic comparatiu dels temps d'execució del mètode de traçat de rectes respecte al de projecció de vòxels acolorits.

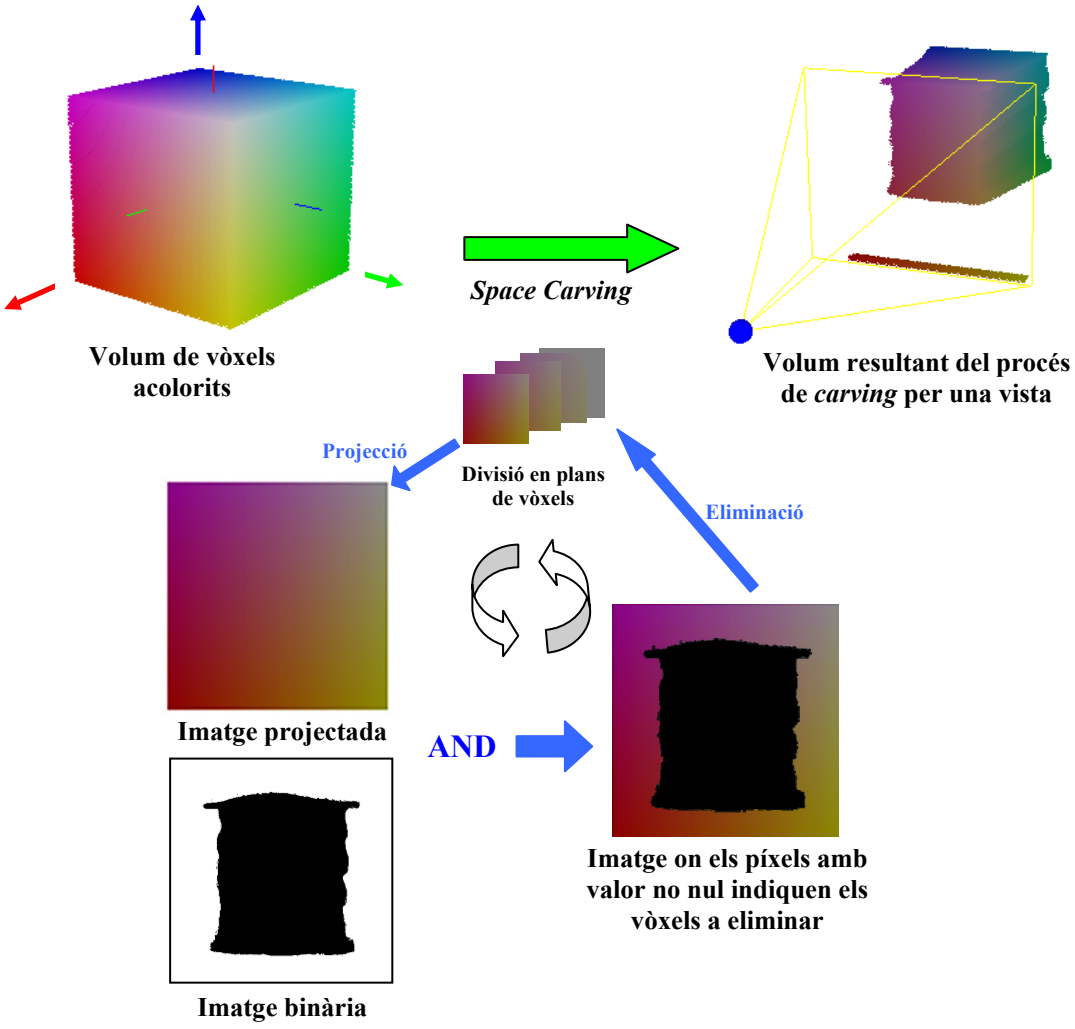


Figura 6.5 Representació esquemàtica del procés de projecció dels vòxels damunt la imatge binaritzada, amb suport del coprocessador gràfic. El resultat de la operació lògica dóna, en els colors, els índexs dels vòxels a eliminar.

6.2.2 Optimització de la projecció, ús d'arbres i mapes de distància.

L'opció plantejada anteriorment vincula el nombre de vòxels disponibles al de colors representables, s'ha estudiat també la optimització de la selecció de vòxels en el cas de representar l'estructura com un *octree*. Aquest mètode és ja clàssic [Connolly 85], que l'aplicava en l'àmbit de cercar la següent millor vista (*Next Best View*), especialment en planificació del moviment de sensors (*sensor planning*). També s'aplica en els mètodes d'extracció de la forma a partir de la silueta *(shape from silhouette*) que està referenciat en l'estat de l'art d'aquesta tesi. El mètode descrit a continuació, es troba dins d'aquesta disciplina, i es pot descriure algorítmicament de la següent manera:

- En primer lloc, cal agafar la imatge, binaritzar-la i crear un mapa de distàncies. Aquest mapa de distància, semblant a la coneguda transformada de distància, farà que cada píxel pugui tenir tres classes de valor: un nombre natural entre u i la resolució de la imatge si és dins de l'objecte o *foreground*, un zero si pertany al contorn de l'objecte, i un nombre negatiu entre menys u i menys la resolució de la imatge si pertany al fons o *background*. Aquest valor codifica com de dins o fora d'un objecte es troba un píxel (veure figura 6.6). Malgrat que la computació estàndard de la transformada distància sols té en compte el cas interior, una petita modificació fent-la alhora per la imatge i la imatge negada en la mateixa iteració dóna el resultat esperat.

-7	-7	-6	-6	-6	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-6	-6
-7	-6	-6	-5	-5	-5	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-5	-5	-6
-6	-6	-5	-5	-4	-4	-4	-3	-3	-3	-3	-3	-3	-3	-3	-3	-4	-4	-5	-6
-5	-5	-5	-4	-4	-3	-3	-3	-2	-2	-2	-2	-2	-2	-2	-3	-3	-4	-5	-5
-5	-4	-4	-4	-3	-3	-2	-2	-2	-1	-1	-1	-1	-1	-2	-2	-3	-4	-4	-5
-4	-4	-3	-3	-3	-2	-2	-1	-1	-1	0	0	0	-1	-1	-2	-3	-3	-4	-5
-4	-3	-3	-2	-2	-2	-1	-1	0	0	0	1	0	0	-1	-2	-2	-3	-4	-5
-4	-3	-2	-2	-1	-1	-1	0	0	1	1	1	1	0	-1	-1	-2	-3	-4	-5
-4	-3	-2	-1	-1	0	0	0	1	1	2	2	1	0	0	-1	-2	-3	-4	-5
-3	-3	-2	-1	0	0	1	1	1	2	2	2	1	1	0	-1	-2	-3	-4	-5
-3	-2	-2	-1	0	1	1	2	2	2	3	2	2	1	0	-1	-2	-3	-4	-5
-2	-2	-1	-1	0	1	2	2	3	3	3	3	2	1	0	-1	-2	-3	-4	-5
-2	-1	-1	0	0	1	2	3	3	3	3	3	2	1	0	-1	-2	-3	-4	-5
-2	-1	0	0	1	1	2	2	2	2	3	2	2	1	0	-1	-2	-3	-4	-5
-2	-1	0	1	1	1	1	1	1	2	2	2	1	1	0	-1	-2	-3	-4	-5
-2	-1	0	0	0	0	0	0	1	1	1	1	1	0	0	-1	-2	-3	-4	-5
-2	-1	-1	0	0	-1	-1	0	0	0	0	0	0	0	-1	-1	-2	-3	-4	-5
-2	-2	-1	-1	-1	-1	-1	-1	0	0	-1	-1	-1	-1	-1	-2	-2	-3	-4	-5
-3	-2	-2	-2	-2	-2	-2	-1	-1	-1	-1	-2	-2	-2	-2	-2	-3	-3	-4	-5
-3	-3	-3	-3	-3	-3	-2	-2	-2	-2	-2	-2	-3	-3	-3	-3	-3	-4	-4	-5

Figura 6.6 Exemple de la transformada de distància amb valors positius i negatius, respecte a la posició d'un objecte segmentat (en blau). Per al càlcul s'ha utilitzat veïnatge-8 en els píxels.

- Començar a dividir l'espai de treball en vuit elements iguals (veure figura 6.7) i de forma recursiva, cada un dels fills resultants vàlids.

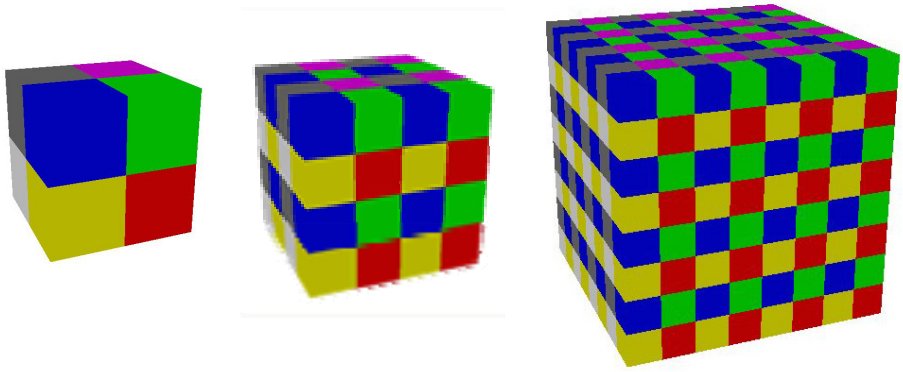


Figura 6.7 Divisió recursiva del volum de vòxels on cada cub es divideix en vuit parts iguals (d'esquerra a dreta). Correspon a la representació d'una estructura d'arbre *octree*.

- Projectar el punt central de cada node fill damunt la imatge amb el mapa de distàncies. Si es projecta damunt un píxel amb valor positiu que guarda un valor superior a la meitat de la mida del que representa el vòxel mantenir, guardar el vòxel sencer i no cal seguir dividint-lo. Si es projecta damunt un píxel amb valor negatiu que en mòdul és major que la meitat de la mida del vòxel eliminar-lo sencer i no cal seguir dividint. En qualsevol altre cas, caldrà continuar fent-hi divisions. (veure figura 6.8). D'aquesta manera es poden ràpidament branques de l'octree representant milers de vòxels.

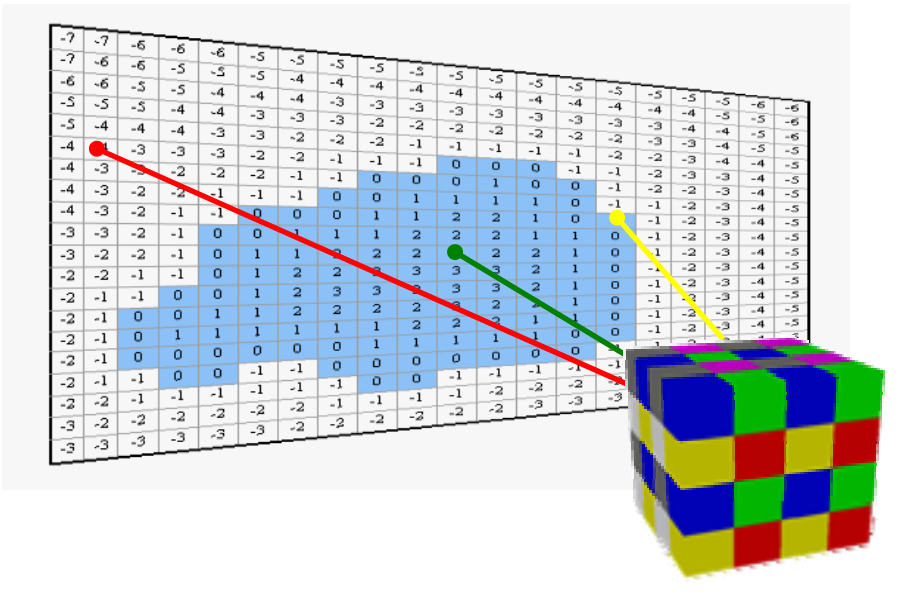


Figura 6.8 Mostra del procés de selecció de vòxels: si es projecta damunt un píxel amb distància positiva major que la mida del vòxel (marcador verd) el vòxel sencer es manté. Si es projecta en un nombre negatiu de mòdul major que la mida, cal eliminar-lo sencer (marcador vermell). En altre cas, s'ha de dividir el vòxel (marcador groc).

- Repetir el procés fins arribar a la profunditat màxima o no quedin més divisions a fer.

Aquest mètode descrit de projecció de l'estructura dels *octree* damunt el mapa de distàncies es pot escriure en llenguatge algorísmic de la següent manera, es suposa que la variable imatge amb el mapa de distàncies és global i que el mètode es pot definir recursivament, tal com surt de forma natural:

```
#definir MIDAMAX VALOR

DefinicioTipus Estructura                    // Tipus necessari pel vòxel
{
        enter     valor;
        Punt3D  pos;
        Punter    enllaç
}
TipusVoxel;

MetodeCarvingArbre( )                        // Aquesta funció que realitza el carving amb arbre
{
        TipusVoxel  V;
        imatge_dist = CalculaMapaDistancia ( );
                                             // Funció que calcula el mapa de distància.
        Dividir ( V, MIDAMAX);
}

Dividir ( TipusVoxel pare, int mida)         // Aquesta és la funció recursiva de divisió del vòxel.
{
        TipusVoxel voxels[8];
        CreaEstructura (pare, voxels[i] ,mida );
                                             // Funció que inicialitza els camps del TipusVoxel
                                             // Valor a -1, enllaç a NULL i pos al lloc geomètric
        pare.enllaç = voxels [i] ;

        /* Procés de space carving */
        per (i = 0; i<8; i++)
        {
                si ( imatge_dist [ Projecció ( imatge_dist , voxels [i].pos )]  < - mida )
                        voxels [i]. valor=0;
                si ( imatge_dist [ Projecció ( imatge_dist , voxels [i].pos )] >  mida )
                        voxels [i]. valor=1;                    // Projecció troba el lloc
                                                                // del vòxel a la imatge
                si ( (voxels[i] !=0) && (voxels[i]!=1)  && (mida>1))
                        Dividir (voxels[i], mida / 2 );
        }
}
```

Considerant que el procés de projecció del centres dels nodes representats a l'arbre pot accelerar-se també amb l'ús de la targeta gràfica, queda un algorisme de cost similar al de la equació 6.2, és a dir de cost cúbic respecte al costat del cub de vòxels, amb una part depenent del processador i una del coprocessador gràfic. La diferència principal és que aquest cost es veu afectat per un factor de probabilitat P:

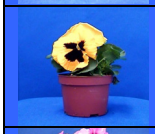
$$\text{Cost(projecció arbre)} = k_0 + P \cdot k_1 \cdot (\text{mida\_V}^2 \cdot \text{mida\_V}) \qquad \text{(Eq. 6.3)}$$

k_0: cost d'inicialització d'estructures.
k_1: cost de la decisió de selecció de vòxel.
mida_V^2: cost de la projecció d'una estructura al coprocessador gràfic.

El factor P és depenent de l'objecte; per un objecte que ocupi tota la imatge o en cas de l'absència de l'objecte amb una iteració es podria decidir que tot el volum ha de quedar-se o ser eliminat. El cas pitjor és aquell en que l'objecte és representat per una graella de píxels de valors alternatius zero / u. En aquest cas l'algoritme té un cost igual al de l'apartat 6.2.1 però amb l'avantatge de poder codificar espais de vòxels de dimensions o resolució sols limitades per la memòria total del sistema (incloent discs). En qualsevol cas l'algoritme presentat de projecció de vòxels representats per arbres tipus *octree* mai tindrà un cost pitjor als anteriors.

6.2.3 Emmagatzemament de la informació en vòxels. Exemples del procés.

S'han experimentat doncs tres maneres d'implementar el mètode d'esculpit de vòxels i gravar la informació al disc: projecció de rectes damunt el volum de vòxels, projectar un volum de vòxels representat com una matriu tridimensional damunt les imatges i projectar els vòxels representats segons l'estructura d'arbre damunt les imatges. La taula 6.1 mostra, per tres objectes diferents, el temps d'execució de cada un d'aquests mètodes i la mida del fitxer necessari per emmagatzemar el volum de vòxels. En els dos casos on s'usa la matriu tridimensional, es mostra el que seria la mida màxima (un *byte* per vòxel) pel mètode de rectes i la llista de vòxels actius de la superfície de l'objecte pel mètode de projecció de vòxels acolorits. Aquesta representació permet reduir molt la mida del fitxer permetent igualment reconstruir l'estructura, És el que s'anomena carcassa envoltant convexa o *convex hull* d'un objecte. En el cas dels arbres *octrees* la representació és, per cada node una taula de vuit nombres que indiquen si hi ha fill o no, i si n'hi ha, apunta a la taula que representa el fill.

	Mètode rectes		**Mètode projecció**		**Mètode arbre *octree***	
Objecte	Temps	Mida	Temps	Mida	Temps	Mida
	101,4s	2MB	30,2s	412KB	12,1s	350KB
	101,4s	2MB	31,7s	421KB	14,2s	360KB
	101,4s	2MB	33,8s	501KB	19,4s	450KB

Taula 6.1 Comparació del temps d'esculpit i de la mida de l'estructura de dades a emmagatzemar amb el mètode de intersecció rectes-vòxels, projecció de vòxels acolorits i projecció de l'estructura en arbre en un espai de 128x128x128 vòxels amb 32 vistes de l'objecte (PC P-IV, 2GHz, 512MB RAM GeForce IV 512MB VRAM, mesures de temps fetes amb el *High Performance Timer* HPT)

La qualitat de la reconstrucció no és depenent del mètode usat ja que només implementen variants en la manera d'executar la projecció dels vòxels i en la mida total de l'emmagatzemament, però treballant sempre amb la mateixa idea de distància entre vòxel i projecció. De la qualitat de la reconstrucció, se'n parlarà en els següents

apartats. La figura 6.9 mostra el resultat de l'aplicació de l'esculpit de vòxels amb dos dels objectes avaluats. La reconstrucció mostrada és en els dos casos pel mètode de projecció del volum de vòxels acolorit i amb una resolució de 128^3 vòxels. Malgrat que per aconseguir una bona qualitat en la reconstrucció cal tenir una resolució de vòxels similar a la resolució en píxels (de l'ordre de 512^3), pel càlcul del mapa dens de disparitat resolucions inferiors donen ja resultats millors que els mètodes d'aparellament estèreo.

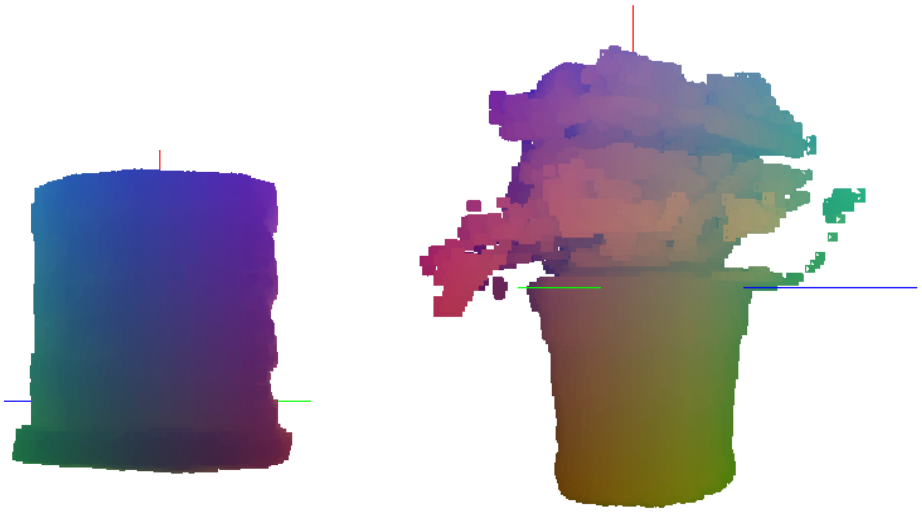


Figura 6.9 Exemple de reconstrucció en vòxels dels objectes urna funerària romana (esquerra) i una planta viola cornuta (dreta) amb una resolució de 128^3 vòxels projectats damunt 32 vistes.

6.3 Relació entre les vistes i els vòxels (1): acceleració del procés de *space carving*.

En aquest apartat es pretén cercar una manera d'accelerar el procés de costós de space carving, aprofitant la informació donada per les pròpies vistes de l'objecte.

6.3.1 Plantejament.

Fins el moment actual, s'ha suposat que el procés d'esculpit de l'espai de vòxels es realitza amb un conjunt donat de vistes. Cal preguntar-se: quin conjunt de vistes serà seleccionat? Una primera opció és usar totes les vistes disponibles. Una altra és cercar un criteri per discriminar quines vistes s'han d'usar i quines no. Donat que en els experiments plantejats s'assumeix que es pot disposar de totes les vistes de l'objecte amb la resolució del posicionador (veure capítol 3), no es planteja un problema de planificació de moviments d'un sensor, sinó simplement de selecció de les vistes per a l'esculpit. Per a fer aquesta selecció s'ha seguit el següent **raonament**:

- El cost del procés és directament proporcional al nombre de vòxels que queden per eliminar.
- En conseqüència seria bo en els primers passos eliminar el màxim nombre de vòxels possible.

- Una vista de l'objecte que, un cop segmentada, té menys àrea que una altra, elimina més vòxels a l'hora de fer l'esculpit.
- Estudiant la funció tridimensional que relaciona l'àrea de l'objecte segmentat amb la posició en longitud i latitud de la vista (veure figura 6.2) es poden trobar les vistes que fan mínims locals en la funció àrea.
- Per tant es pot establir un ordre: executar el procés triant les vistes d'àrea mínima ordenades de menor a major (veure figura 6.10), que serà la manera més ràpida d'eliminar els vòxels .
- Quant més tard es talli aquest procés (en el límit es podria arribar a usar totes les vistes amb un mínim a la funció àrea), més qualitat es tindrà.

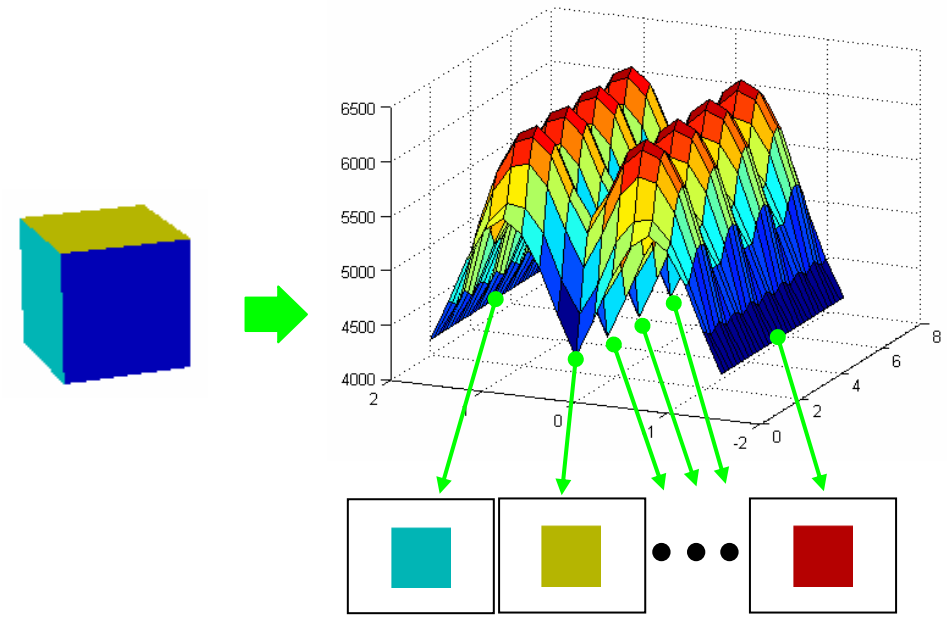


Figura 6.10 Obtenció del conjunt de sis vistes amb mínims a la funció àrea per un objecte cúbic.

Segons aquest raonament, es planteja triar les vistes per al procés d'esculpit dels vòxels seguint l'algoritme de "**esculpit amb vistes ordenades per mínims en la funció àrea**" plantejat a continuació:

1) Calcular la funció àrea de l'objecte segmentat respecte la posició de totes les vistes.
2) Cercar-hi tots els mínims locals i ordenar-los pel seu valor, de menor a major.
3) Passar en aquests ordre les vistes a qualsevol dels mètodes de *space carving* exposats.

6.3.2 Anàlisi del rendiment obtingut.

Un cop vist l'algoritme plantejat cal avaluar quina precisió s'assoleix. És evident que a mida que es deixa avançar més l'algoritme l'error disminueix. En aquest punt l'algoritme ofereix paral·lelismes amb els compressors estàndards d'imatge, que a mida que es deixen passar més components milloren la qualitat; aquí, a mida que s'utilitzen més vistes amb mínims, i per tant es dedica més temps de processador, la qualitat millora. Donat que amb mètodes alternatius com el del làser plantejat al capítol cinquè,

o el d'esculpit de vòxels usant totes les vistes de manera estàndard ja s'obté una reconstrucció de l'objecte, es té la possibilitat d'establir comparacions entre el volum reconstruït pel mètode de vistes ordenades i els altres. D'aquesta manera s'ha pogut generar la gràfica mostrada a la figura 6.11, on es veu que el mètode de esculpit per vistes d'àrea mínima arriba més ràpidament al volum límit dels mètodes de reconstrucció d'objectes. La diferència entre el volum real i el límit del mètode és degut a les possibles concavitats de l'objecte, que un mètode com el del làser si que pot trobar i el *carving* no.

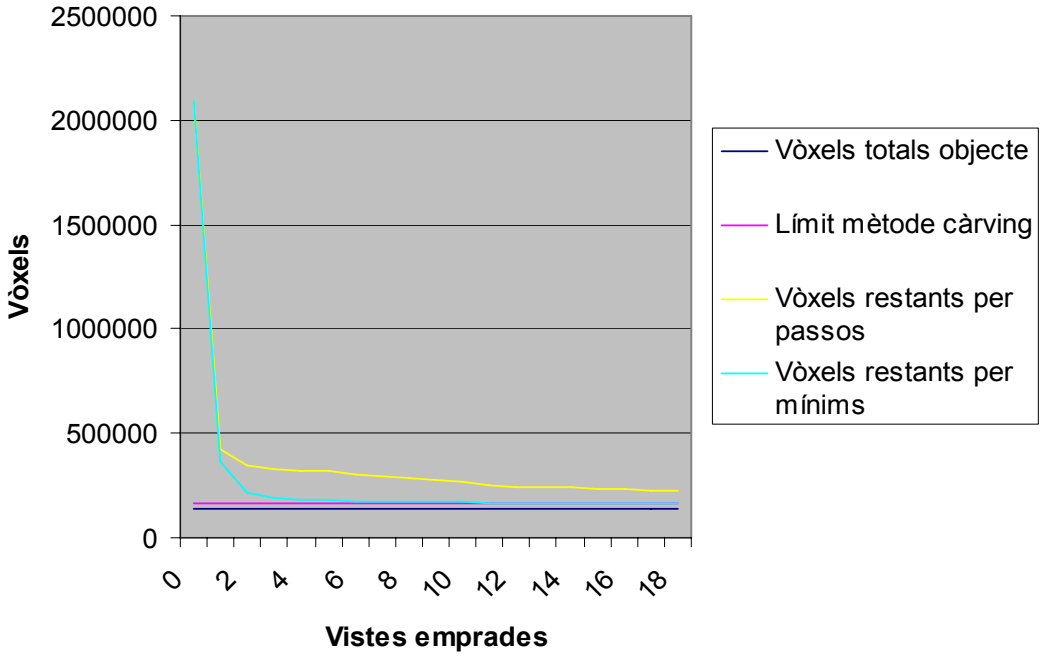


Figura 6.11 Mostra de la convergència del nombre de vòxels restants al volum segons el mètode d'ordenació de mínims (blau) i un de distribució uniforme de vistes (groc), per l'objecte urna funerària. S'han tingut resultats semblants pels altres objectes.

Existeixen casos (cub, esfera) de que el mètode pot necessitar ésser complementat. En el cas del cub (si fos perfecte) sols hi hauria sis vistes amb mínims, per guanyar més precisió caldria seguir amb d'altres. En el cas de l'esfera totes tenen el mateix valor en la funció àrea. De totes maneres, els experiments fan pensar que és el que més ràpid s'aproxima a la solució. Com que es creu que l'anàlisi de la funció àrea és una bona eina, es planteja la següent conjectura. Llur demostració es deixa per treballs futurs que puguin sortir d'aquesta tesi:

Conjectura 1: "El mètode d'esculpit de l'espai de vòxels per vistes ordenades segons els mínims en la funció àrea, és el que més ràpid convergeix cap a la solució."

6.4 Relació entre les vistes i els vòxels (2): Cerca de les vistes mínimes per la descripció d'un objecte.

De la mateixa manera que estudiant la funció àrea definida damunt l'espai de vistes es pot trobar una estratègia per la selecció de les vistes emprades en el procés d'esculpit d'imatges, es vol utilitzar aquesta funció per trobar un conjunt de vistes que,

combinada amb el procés de síntesi descrit al capítol quart, pugui generar qualsevol altre vista d'un objecte.

Com s'ha vist en la descripció del mètode de síntesi de vistes, és necessari tenir un parell de vistes i el mapa de disparitat que les relaciona. Per poder trobar el mapa de disparitat i que aquest tingui sentit, és necessari que entre les dues vistes s'estableixi l'anomenada regió estèreo, és a dir, que una part de la superfície de l'objecte sigui projectada a les dues vistes alhora. Això implica que les vistes han d'estar relativament properes entre elles.

Per altra banda, com major sigui el nombre de vistes i mapes de disparitat emmagatzemats, major serà l'espai a disc necessari i serà més difícil poder usar el mètode de síntesi de vistes en aplicacions de representació a distància o en aplicacions de realitat augmentada que vulguin gestionar diversos objectes. En conseqüència, serà bo trobar un criteri per determinar el conjunt de vistes que millor satisfaci el compromís entre qualitat de les imatges sintètiques i quantitat d'informació a enregistrar.

Experimentalment, exceptuant el cas singular caracteritzat per l'esfera o regions esfèriques, on totes les vistes tenen la mateixa àrea, es pot afirmar que:

- Donat que (òbviament) les vistes on la funció àrea determina un màxim representa un punt on una gran part de la superfície de l'objecte és projectada.
- Entre dues vistes amb àrea màxima veïnes existeix una regió estèreo que permetrà calcular el mapa de disparitat pels punts d'aquesta regió.
- Per poder sintetitzar vistes de tots els punts d'una superfície cal que, com a mínim cada punt hagi estat projectat en com a mínim dues vistes.

Tenint en compte tot això, es pot determinar que el criteri de selecció de vistes a emprar per trobar un conjunt suficientment representatiu d'un objecte pot ser triat entre aquests tres:

Criteris de selecció de vistes per a síntesi de vistes.

1) Per objectes tipus esfera, o que tenen grans porcions de superfície que projecten la mateixa àrea, o per objectes dels que no s'ha volgut/pogut extreure la característica àrea al capturar les vistes, utilitzar una distribució uniforme de vistes sobre l'objecte (per exemple, una cada 15 graus, una cada 30 graus, etc.).

2) Per un objecte del qual s'ha reconstruït l'estructura en vòxels completa, es pot plantejar la relació vòxels - vistes tal com mostra la taula 6.2. En aquest cas es mostra només aquesta relació per una direcció del recorregut de les vistes; es podria entendre de manera que cobrís les dues direccions resultant una taula tridimensional. Assumint que els vòxels són representatius del conjunt de punts de la superfície de l'objecte (en el límit matemàtic, suposant que la mida del vòxel tendeix a zero segur que ho és) i que cal que un punt sigui vist com a mínim dos cops per a poder calcular els mapes de disparitat i sintetitzar vistes, l'anàlisi d'aquesta taula vòxels - vistes ens donarà el conjunt mínim de vistes que veuen tots els vòxels com a mínim dues

vegades. L'algoritme a aplicar damunt la taula 6.2 serà: trobar el nombre mínim de columnes que garanteix que cada fila apareix més d'una vegada en una vista.

Vòx\vista	0º	5º	10º	15º	20º	25º	30º	35º	40º	45º	50º	...	355º
1	S	N	N	N	N	N	N	N	N	N	N	N	S
2	S	S	N	N	N	N	N	N	N	N	N	N	S
3	S	S	S	S	N	N	N	N	N	N	N	N	S
4	S	S	S	S	S	N	N	N	N	N	N	N	N
5	S	S	S	S	S	N	N	N	N	N	N	N	N
6	N	N	S	S	S	S	S	N	N	N	N	N	N
7	N	N	S	S	S	S	S	N	N	N	N	N	N
8	N	N	N	S	S	S	S	N	N	N	N	N	N
9	N	N	N	N	S	S	S	N	N	N	N	N	N
10	N	N	N	N	N	S	S	S	N	N	N	N	S
11	N	N	N	N	N	N	S	S	N	N	N	N	S
12	N	N	N	N	N	N	S	S	S	N	N	N	S
13	N	N	N	N	N	N	N	S	S	S	N	N	S
14	N	N	N	N	N	N	N	S	S	S	N	N	N
15	N	N	N	N	N	N	N	N	S	S	S	N	N
16	N	N	N	S	S	S	S	N	S	S	S	N	N
17	N	N	N	S	S	S	S	S	N	S	S	N	N
...	...	...	...	...	...	...	...	...	...	...	...	...	...
MAX	N	N	N	N	N	N	N	N	N	N	N	N	S

Taula 6.2 Exemple de la taula que vincula totes les vistes disponibles amb tots els vòxels representats. Si es troba el mínim nombre de columnes que contempla cada vòxel com a mínim dos cops, es podran calcular els mapes de disparitat que permetran interpolar vistes intermèdies. S'han marcat en color verd 4 columnes que per l'exemple compleixen la condició.

3) Tenint calculada la funció àrea de l'objecte segmentat per totes les vistes, i havent determinat experimentalment que el conjunt de vistes amb màxims locals en la funció àrea representa de manera vasta l'objecte i que permet calcular els mapes de disparitat, es conjectura que:

> **Conjectura 2:** "el conjunt de vistes on la funció àrea de l'objecte segmentat té màxims locals és un bon subconjunt de vistes per usar en el procés de síntesi de vistes de l'objecte".

La figura 6.12 mostra com s'ha emprat aquest criteri als experiments per trobar un conjunt mínim de vistes representatives. En el cas del cub es troben vuit vistes amb màxims locals que permeten trobar els mapes de disparitat i interpolar totes les vistes de l'objecte.

Com en el cas de la primera conjectura exposada, existeixen casos, com el d'una esfera, en que no es poden determinar màxims en la funció àrea de l'objecte. En aquest cas caldria aplicar un dels criteris de selecció anteriors: distribució uniforme de vistes damunt l'objecte o cerca del nombre mínim de vistes que situen tots els vòxels de la superfície de l'objecte en una regió estèreo. Una demostració formal de la certesa o falsedat d'aquesta afirmació, generalitzada a qualsevol objecte, es deixa per a treballs futurs.

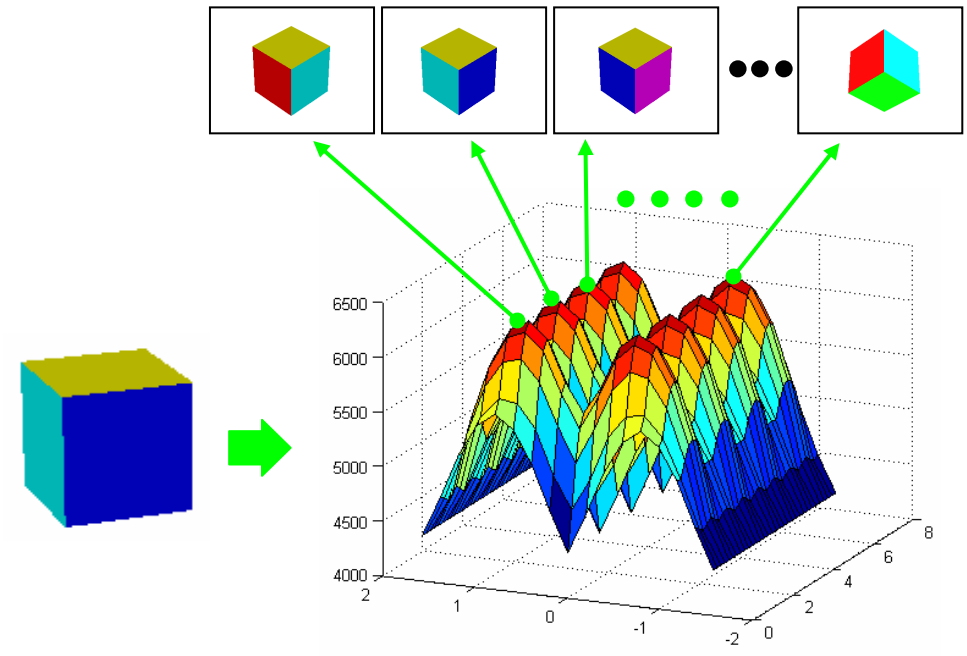


Figura 6.12 Obtenció del conjunt de vuit vistes amb màxims a la funció àrea per un objecte cúbic. Amb aquestes vuit vistes i els seus mapes de disparitat, hi haurà prou per a interpolar-ne qualsevol altre.

6.5 Exemple d'obtenció de vistes amb el mètode de selecció i síntesi per interpolació.

Un cop trobat un criteri per la selecció d'un conjunt representatiu de vistes de l'objecte ja es podrà aplicar el mètode de síntesi de vistes per interpolació mostrat al capítol quart d'aquesta tesi. A tall d'exemple, es mostra ara el resultat de l'aplicació de la selecció de vistes i interpolació amb un objecte capturat pel sistema d'adquisició. La figura 6.13 mostra a la part superior la funció àrea de l'objecte segmentat per una volta damunt l'objecte. D'aquesta sèrie de valors s'han seleccionat les vistes amb àrea mínima per al procés de reconstrucció tridimensional via *space carving* explicat anteriorment. També s'han seleccionat dues vistes amb màxims a la funció àrea amb les que s'ha procedit a aplicar el mètode de síntesi de vistes desenvolupat al capítol quart, rectificant les imatges, obtenint el mapa de disparitat i finalment, interpolant la nova vista de l'objecte.

En el capítol de resultats d'aquesta tesi es parlarà de la qualitat de la imatge sintetitzada, de quins criteris s'han aplicat per a mesurar aquesta qualitat, del temps d'execució del procés i com a prop s'està d'arribar a l'anomenat *video-rate*, dels recursos del maquinari emprats pel procés, etc. També es definirà el mètode complet de selecció i síntesi de vistes i es compararà amb els altres mètodes d'obtenció de vistes identificats.

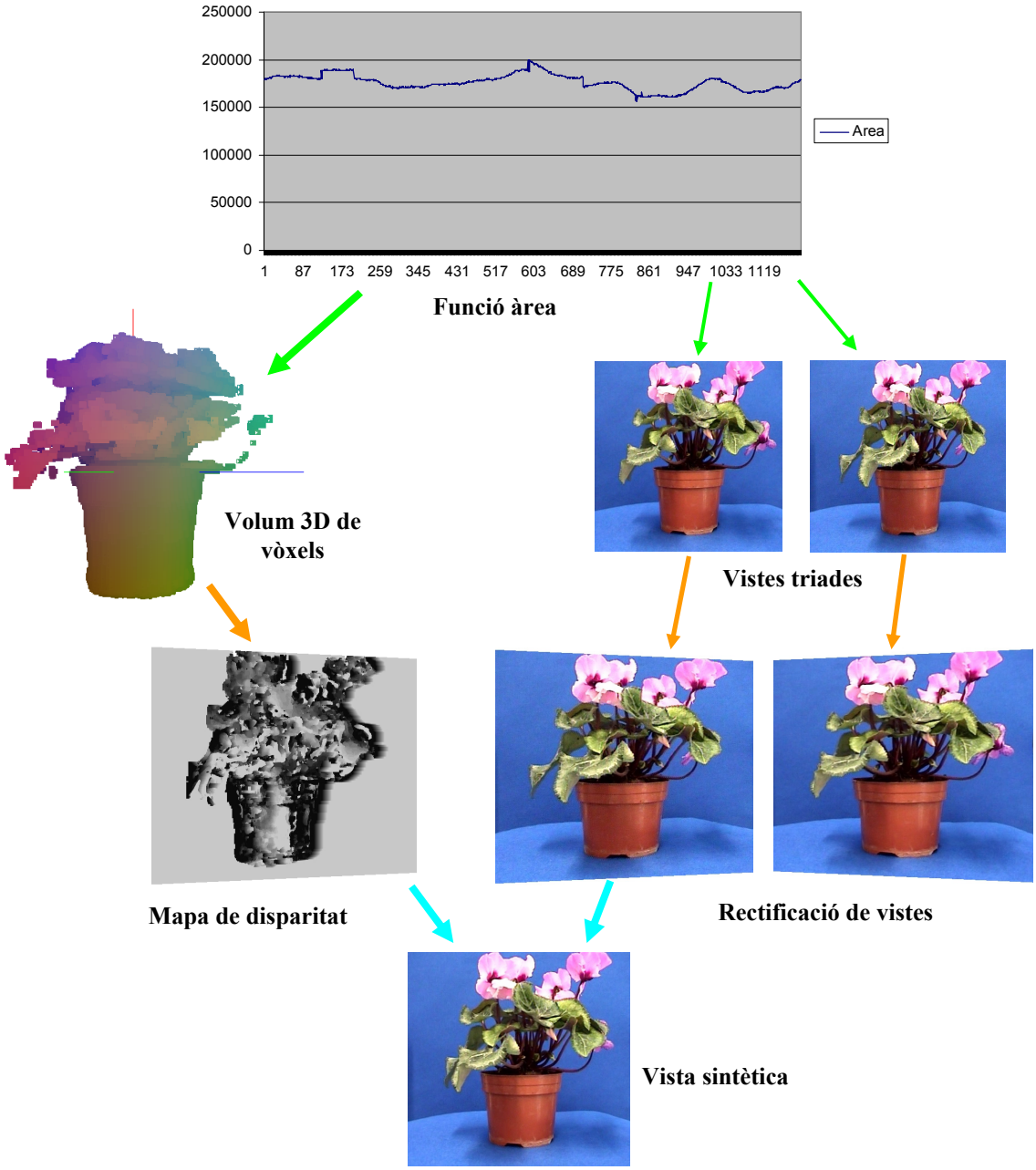


Figura 6.13 Exemple de la selecció i interpolació de vistes amb els passos: anàlisi de la funció àrea, reconstrucció tridimensional, selecció de vistes, rectificació de les vistes, càlcul del mapa de disparitat a partir de la informació tridimensional i síntesi de la nova vista.

6.6 Refinament dels models tridimensionals emprant la síntesi de vistes.

S'ha definit un mètode d'obtenció de vistes sintètiques la qualitat de les quals té, inevitablement, una clara dependència de la bondat de la informació tridimensional obtinguda a partir de les vistes originals. Malgrat haver experimentat diversos mètodes per la recuperació del model tridimensional de l'escena, s'ha preferit treballar amb el mètode d'esculpit del volum de vòxels ja que pot obtenir un mapa de disparitat més dens que un mètode estèreo i és més fàcil de portar a qualsevol entorn i tipus d'objecte

que el de triangulació amb el làser. De totes maneres, com s'ha vist a la figura 6.11, existeix un límit en la qualitat de la reconstrucció assolible amb el mètode d'esculpit de vòxels. Les limitacions d'aquest mètode estan donades per la seva impossibilitat de detectar les concavitats existents en un objecte, tal com posa mostra la figura 6.14.

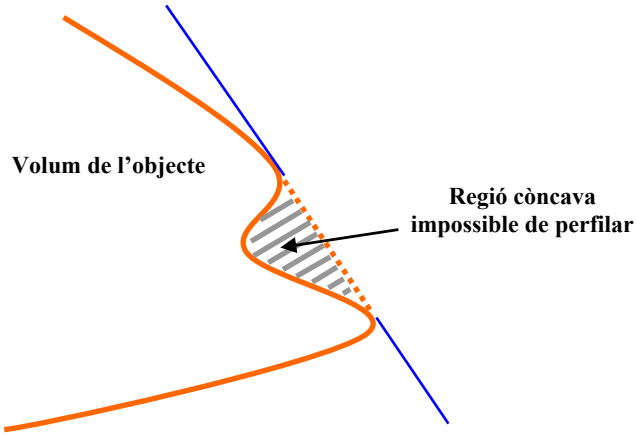


Figura 6.14 Limitació del mètode de *space carving* en la reconstrucció dels volums amb concavitats.

Al fer us de la informació tridimensional per la interpolació de noves vistes, es va detectar que els errors donats per aquestes concavitats en l'objecte es manifesten en les imatges sintètiques. Com que en els experiments s'havien adquirit totes les vistes de l'objecte (veure capítol tercer), es pot comparar la vista interpolada per un punt amb la que s'obté en el sistema real. Aquesta possibilitat d'usar el mètode de síntesi de vistes per millorar la reconstrucció tridimensional que ell mateix utilitza planteja un procés realimentat de refinament del model tridimensional (aportació presentada a [Martin-Aranda 03]). El procés de *space carving* dóna un model bast de l'objecte que convergeix cap a la carcassa envoltant convexa, i en les regions còncaves es pot aplicar un refinament per estereovisió. La determinació d'uns pocs punts mitjançant l'aparellament estèreo, permetrà millorar considerablement l'aproximació dels vòxels al volum de l'objecte, tal com mostra la figura 6.15, on la utilització de dues vistes que identifiquen i triangulen la posició d'un punt dins la regió còncava redueix dràsticament l'error en la reconstrucció.

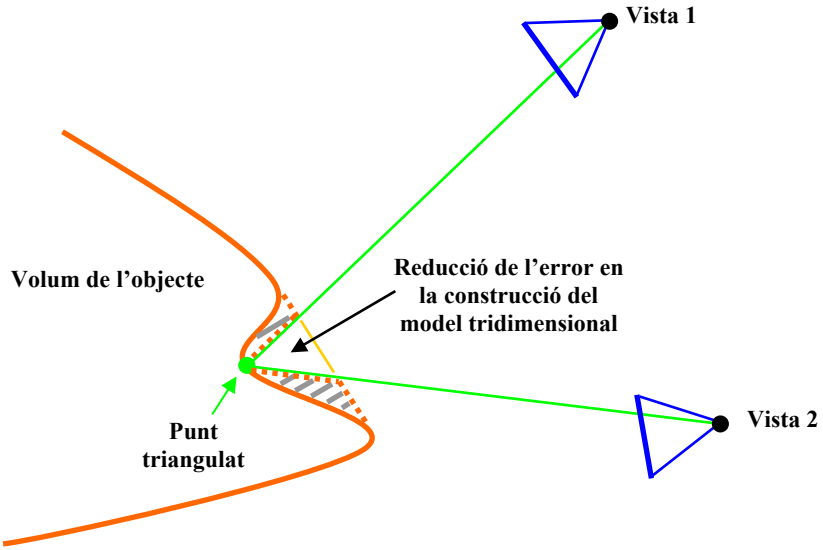


Figura 6.15 Correcció del model tridimensional esculpit amb la informació aportada per una triangulació estèreo.

Així doncs, com a resultat col·lateral del mètode de síntesi de vistes, es pot definir un mètode de **reconstrucció tridimensional d'objectes per esculpit de vòxels refinat amb estereovisió**. L'esquema de la figura 6.16 mostra el procés a aplicar en aquest mètode on, el procés d'aparellament estèreo s'aplica puntualment en les regions on es detecta un error major d'un llindar. Com s'ha dit, l'avaluació dels errors produïts en l'obtenció de les imatges, mereixerà un capítol sencer en la tesi. Alguna de les mesures proposades allà, o una combinació d'elles servirà per la determinació d'aquesta variable error, en els exemples mostrats s'ha emprat una comparació d'imatges pel mètode de suma de diferències al quadrat.

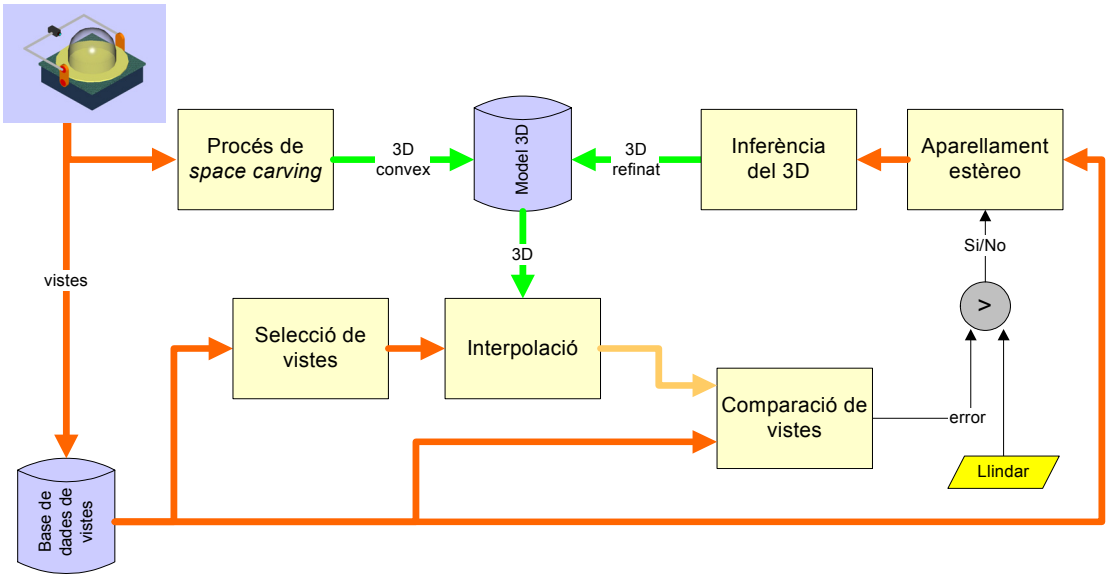


Figura 6.16 Procés de refinament de l'estructura tridimensional obtinguda pel mètode de *space carving* mitjançant la detecció dels errors en la interpolació i correcció mitjançant aparellament estèreo.

L'aplicació del refinament amb estereovisió permet que, en el cas de l'objecte urna funerària mostrat a la figura 6.17, el volum de la reconstrucció amb vòxels s'apropi més al volum suposat de l'objecte, mesurat amb la triangulació làser. Això posa de manifest la bondat d'aquest procés tant en la recuperació de l'estructura tridimensional com en la reducció de l'error provocat per les concavitats en el procés de síntesi de vistes.

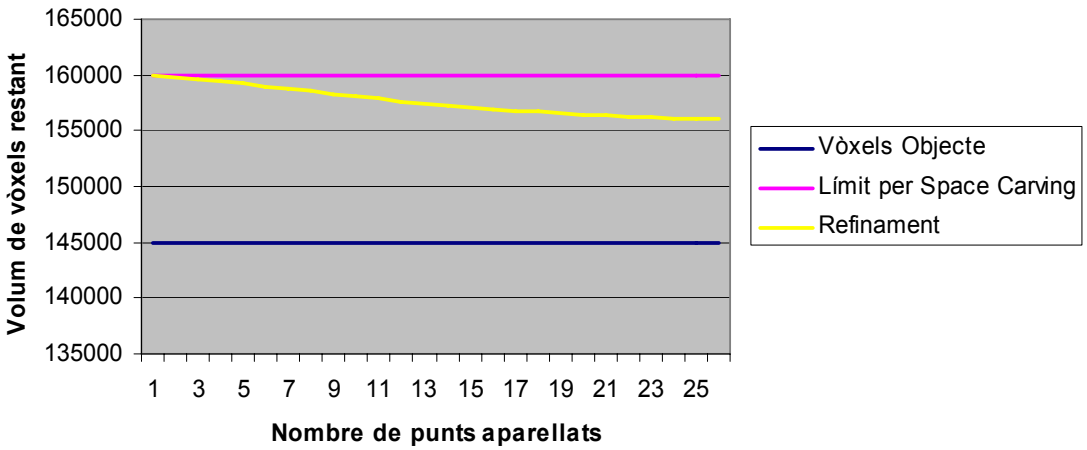


Figura 6.17 Experimentació del refinament del volum de vòxels obtingut mitjançant *space carving* amb estereovisió.

7. Resum dels mètodes d'obtenció de vistes trobats

En aquest capítol es mostrarà el resultat dels estudis presentats fins el moment. S'ha generat una base de dades de vistes per a fer els experiments, i aquesta base de dades s'ha enregistrat en forma de fitxer de vídeo on els índexs de les imatges codifiquen la posició des de les que han estat preses. Després s'ha mostrat un mètode de síntesi de vistes que, a partir d'unes poques seleccionades i informació tridimensional pot generar qualsevol altre. S'ha presentat el mètode d'esculpit de vòxels que obté una bona reconstrucció tridimensional dels objectes i finalment, s'ha cercat la manera d'optimitzar aquesta reconstrucció i cercar el conjunt mínim de vistes que representi l'objecte. Així doncs, com es pretenia, s'ha trobat una via per l'obtenció de totes les vistes d'un objecte a partir d'una selecció de vistes claus i un mètode d'interpolació. En el camí però, s'han identificat dos altres mètodes que també poden donar totes les vistes d'un objecte si hi ha condicions adequades. El primer consisteix en, simplement, usar la base de dades creada per l'experiment per accedir a les vistes de l'objecte i pintar-les. El segon en agafar la informació tridimensional obtinguda, donar-li textura i projectar-la emprant el coprocessador gràfic. A continuació es descriuran amb detall els tres mètodes d'obtenció de vistes identificats.

7.1. Accés a fitxers de vídeo: el primer mètode d'obtenció de vistes.

En el capítol tercer s'ha vist com es poden capturar i guardar les imatges obtingudes, sols queda mostrar el procediment que es pot emprar per a recuperar-les. S'ha explicat que les dades queden emmagatzemades en un fitxer de vídeo a disc, el qual és una seqüència indexada com una pel·lícula. Per la construcció del fitxer, els índexs representen la posició de la qual s'ha pres una vista determinada. Així doncs, si s'han pres M vistes per volta, en un total de N voltes, a diferents alçades consecutives, es pot definir la variable $índex = (M \cdot j + i)$, que representa la longitud i-èssima de la j-èssima latitud capturada.

Quan un sistema de realitat augmentada o de pintat d'objectes a distància requereixi la vista tinguda des d'un punt P, que està observant l'objecte situat en un punt C de l'espai, caldrà seguir la seqüència d'operacions següent per obtenir la imatge (*frame* en el llenguatge dels fitxers de vídeo) guardada en la pel·lícula:

1) Obtenir la distància *d* a la que s'està observant l'objecte.
2) Calcular els angles corresponents a la longitud i latitud observades.
3) Indexar el fitxer amb els dos valors d'angle obtinguts, usant una funció d'accés directe al *frame* (les llibreries estàndard disposen habitualment de la funció *seek(frame)*).
4) Descomprimir, si s'escau, la imatge extreta de la pel·lícula.
5) Aplicar un factor d'escala a la imatge en funció de *d* i de la distància càmera objecte del moment de la captura.
6) Dibuixar la imatge en el sistema.

La figura 7.1 mostra aquest procés d'accés a les vistes. Es la manera més senzilla de mostrar remotament informació fotomètrica preobtinguda d'un objecte.

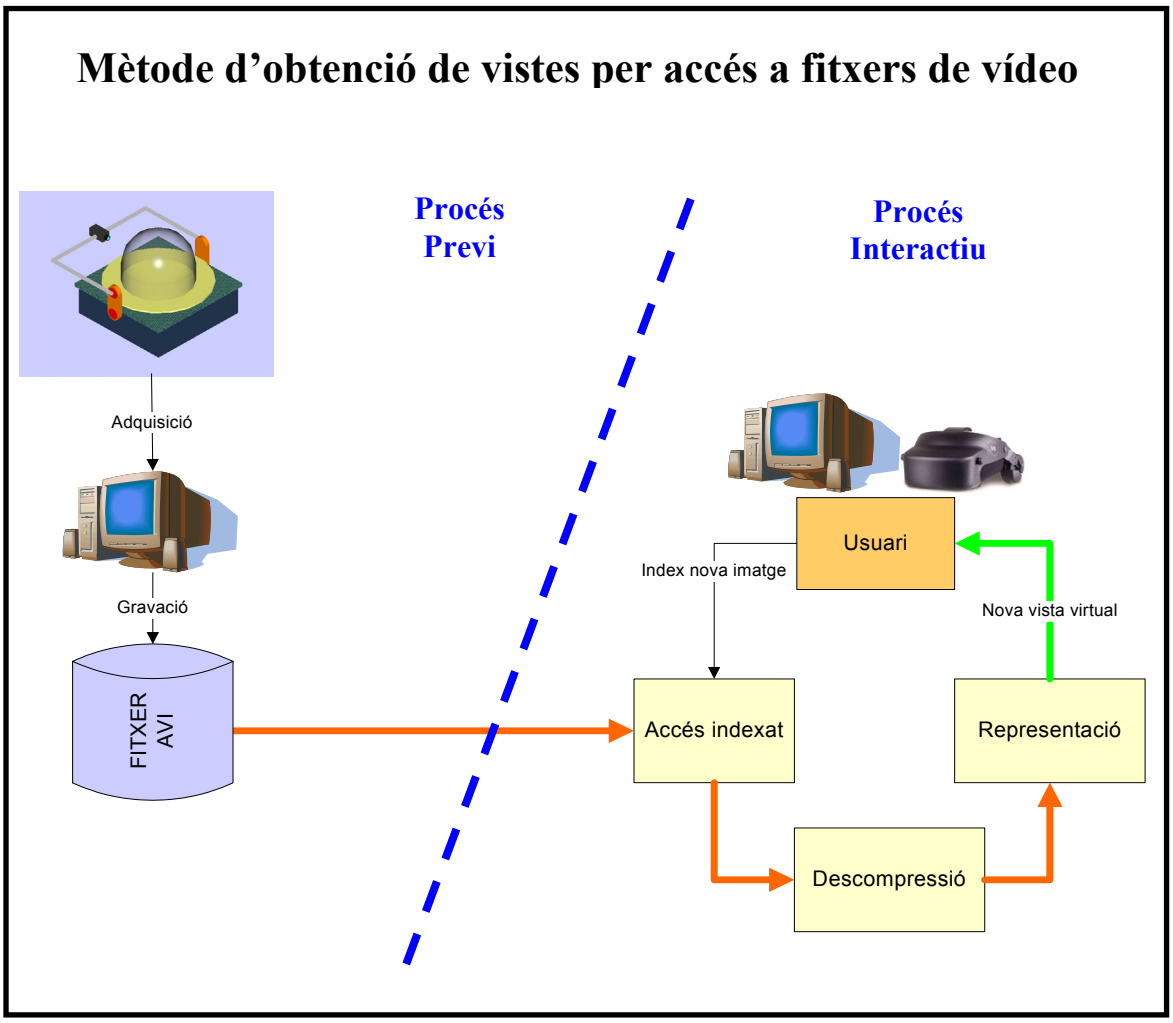


Figura 7.1 Procés d'obtenció de vistes per accés directe a fitxers de vídeo.

Els punts crítics d'aquest mètode seran la mida del fitxer que emmagatzema la seqüència d'imatges, la relació entre aquesta mida i la qualitat de les imatges descomprimides, el nombre necessari d'imatges a guardar per tenir prou resolució, el temps d'accés a disc i la velocitat de descompressió. Les evidents avantatges són la senzillesa i el baix cost en processat. Al capítol de resultats d'aquesta tesi es podran veure les dades concretes de temps d'accés i representació, qualitat obtinguda i comparar-les amb els altres mètodes.

7.2. Representació de models. El segon mètode d'obtenció de vistes.

En la cerca de l'estructura tridimensional de l'escena per obtenir els mapes de disparitat, s'ha evidenciat que un altre mètode per generar les vistes necessàries per aplicacions de realitat augmentada o telepresència consisteix en explicitar l'estructura tridimensional de l'objecte i texturitzar-la amb la informació fotomètrica treta de les vistes. Les targes gràfiques actuals ofereixen un gran suport per aquesta mena de representació (un coprocessador gràfic estàndard va equipat amb una CPU de 500 MHz de rellotge amb 256MB de RAM) i existeixen llenguatges com *OpenGL* i *DirectX* per la descripció de l'estructura i els colors. Així doncs, podem definir un mètode de representació de models tridimensionals tal com mostra la figura 7.2.

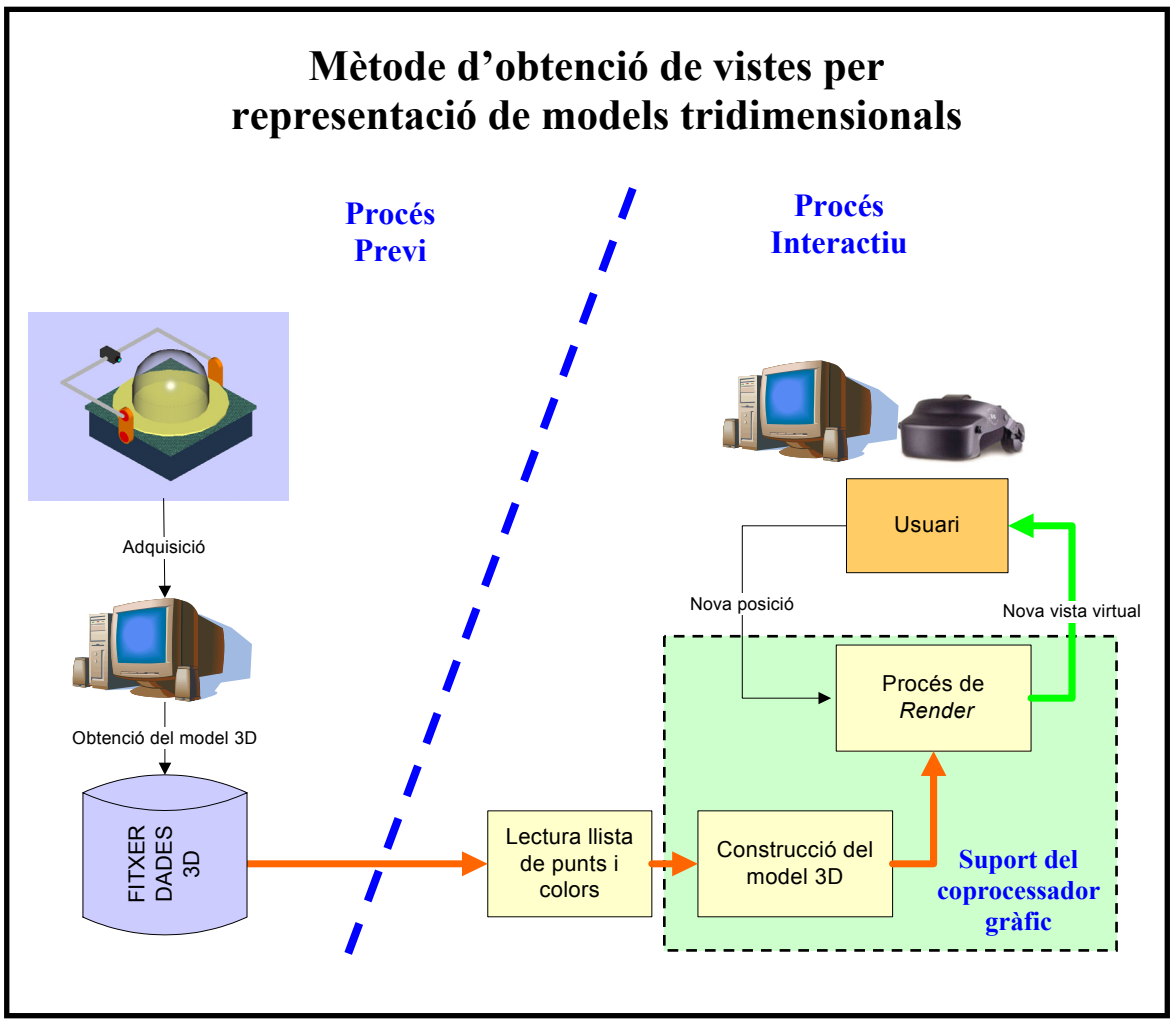


Figura 7.2 Procés d'obtenció de vistes per projecció de models tridimensionals texturitzats, amb suport del coprocessador gràfic (GPU).

La part interactiva del mètode tindrà un primer pas de lectura de la llista de punts i colors d'un fitxer a disc. Aquesta informació serà portada després al coprocessador gràfic on es crearà, per cada triada de punts veïns un triangle al qual s'aplicarà el color corresponent. Aquesta llista de triangles quedarà ja a la memòria del coprocessador gràfic. A partir d'aquest moment sols queda demanar al maquinari les projeccions de la

llista de triangles representativa del volum a mida que l'usuari requereix les noves vistes. Al capítol de resultats d'aquesta tesi s'avalua el rendiment d'aquest mètode i se'l compara amb els altres dos trobats.

7.3 Selecció i síntesi de vistes: el tercer mètode d'obtenció de vistes.

En el capítol quart s'ha exposat un mètode de síntesi de noves vistes a partir d'un parell de vistes d'un objecte, el mapa de disparitat i informació geomètrica de la ubicació de les càmeres en l'espai. En el capítol cinquè i sisè s'ha vist com es pot obtenir el mapa de disparitat i la seva relació amb la informació tridimensional obtinguda de l'objecte. També s'han exposat uns criteris de selecció de vistes que permetran, a partir del conjunt de vistes de l'objecte, seleccionar-ne un subconjunt per ser usades en la síntesi de vista de manera que es pugui obtenir qualsevol altre vista de l'objecte. De la conjunció de tots aquests elements s'obté el que serà tercer mètode d'obtenció de vistes d'un objecte per les aplicacions de realitat augmentada o telepresència proposades.

Es proposa així el **mètode de selecció i síntesi de vistes**. Aquest mètode, com els altres consistirà en dues fases; una prèvia que es pot fer *offline* en la que no tindrà massa importància la utilització de recursos com ara temps de processador i espai a disc i una fase *online* o interactiva en la que l'usuari anirà demanant vistes de l'objecte i en que serà clau garantir un temps d'execució de l'ordre del temps de refresc d'imatge en els dispositius de vídeo. Entre una fase i l'altra queda informació final gravada a disc. Quant menor sigui aquesta informació més portable serà el conjunt de dades, és a dir, la possibilitat de representar vistes d'objectes en entorns de realitat augmentada o telepresència. Vist en conjunt, el mètode consisteix en els següents passos:

Procés previ:

1. Adquirir tantes vistes com sigui possible de l'objecte a representar, s'hauran de conèixer els paràmetres de calibració de la càmera.
2. Triar un conjunt de vistes representatiu de l'objecte, segons algun d'aquests criteris:
   - Si es disposa d'un nombre reduït de vistes, usar-les totes.
   - Si no es té cap informació a priori de l'objecte, usar una distribució uniforme de vistes.
   - Si es disposa d'informació sobre la segmentació de les vistes de l'objecte, extreure la característica àrea i usar-la per determinar vistes singulars (projeccions màximes i mínimes).
   - Si es disposa d'informació tridimensional de l'objecte, cercar el conjunt mínim de vistes que garanteix que cada punt de la superfície ha estat vist com a mínim dues vegades per poder tenir els mapes de disparitat.
3. Per les parelles de vistes properes, calcular els mapes de correspondència entre píxels, usant alguna de les tècniques descrites:
   - Correspondència estèreo, amb un mètode com el de cerca del punt més semblant minimitzant la SSD.
   - Triangulació per làser dels punts de la superfície i projecció a les vistes per obtenir els punts corresponents.

- Reconstrucció del volum de vòxels per *space carving* i projecció a les vistes obtenint les coordenades dels punts corresponents.

4. Guardar aquesta informació a la base de dades del disc: vistes, posició de la càmera i mapes de correspondència.

Per la part del procés previ es mostren tres de les opcions més fàcils d'implementar a la figura 7.3., ja que l'ús de la triangulació per làser s'ha considerat una ajuda puntual per la comparació de la precisió dels diferents mètodes però es voldria evitar en el plantejament final. Així doncs es mostren les opcions de reconstrucció per *space carving* i estereovisió. En la selecció de vistes es presenta l'opció estudiada en aquesta tesi d'utilització de la funció d'àrea de l'objecte segmentat per determinar-les o la solució genèrica de mostrejar l'espai de vistes uniformement.

Selecció i síntesi de vistes: opcions pel procés previ

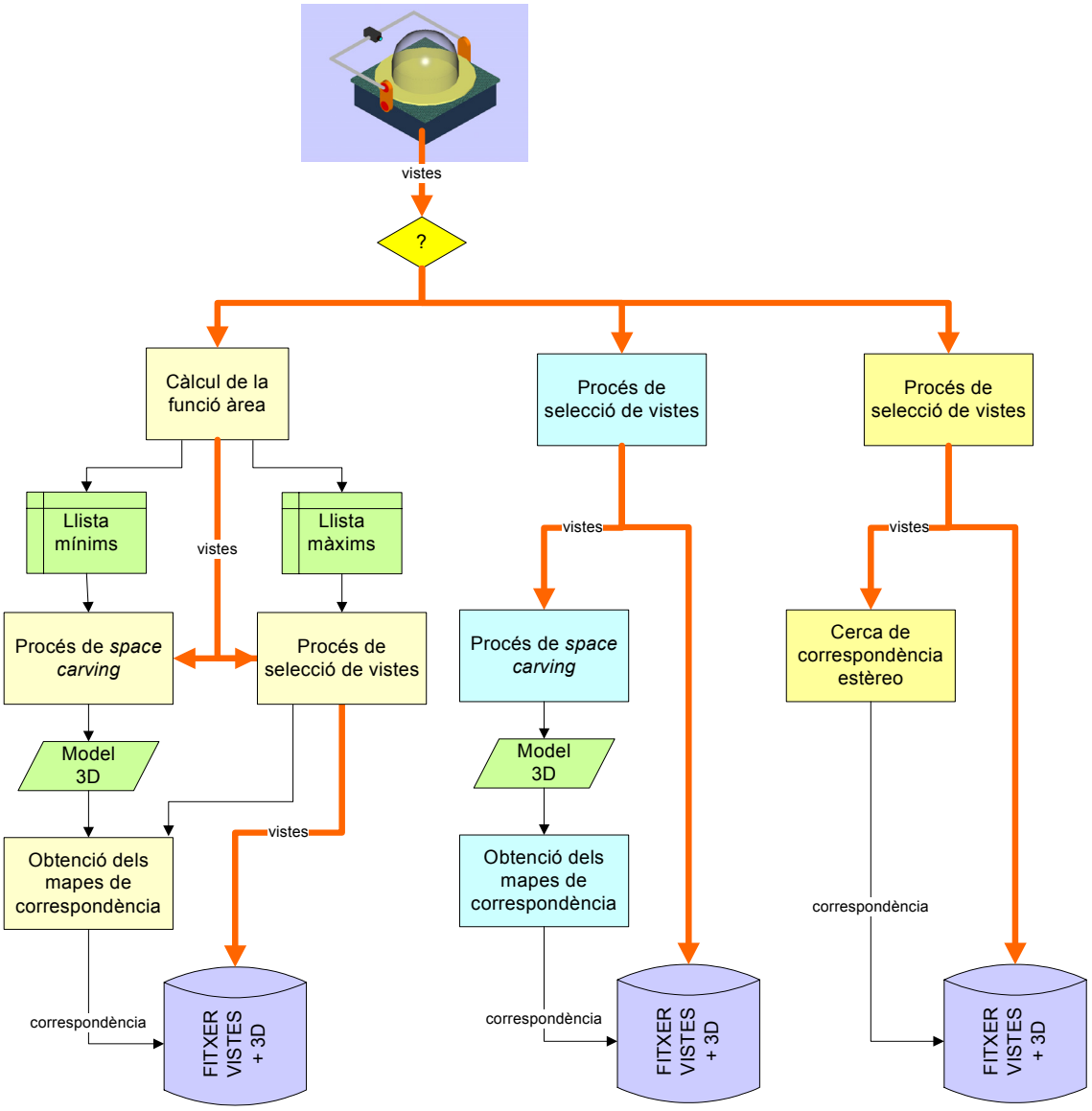


Figura 7.3 Plantejament de les tres opcions principals en el procés previ d'adquisició de la informació: la primera implica la selecció de vistes pel criteri d'àrea i la reconstrucció tridimensional per *space carving*. La segona una selecció de vistes per mostreig i una recuperació del 3D per *carving* també. La darrera proposa una selecció de vistes per mostreig i de la informació de correspondència per mètodes d'estereovisió.

Procés interactiu:

1. Obtenir la posició de la qual es vol obtenir la nova vista.
2. Cercar les dues vistes gravades més properes a la desitjada i construir el pla de reprojecció amb les posicions de les tres càmeres, dues corresponents a les vistes preobtingudes i una a la que serà vista virtual.
3. Rectificar les imatges projectant-les al pla de manera que s'obtinguin les condicions de geometria epipolar.
4. Convertir la informació de correspondència entre punts a disparitat segons la geometria obtinguda.
5. Aplicar el mètode d'interpolació descrit al capítol quart.
6. Representar la imatge en el dispositiu de sortida de forma lliure o registrada en un model tridimensional.

En qualsevol cas, després del plantejat i experimentat en aquesta tesi, a l'hora de definir el procés en conjunt, es prefereix l'opció d'obtenir la informació tridimensional via *space carving* ja que dona un mapa de correspondència i disparitat molt més dens que el mètode d'estereovisió. Les vistes triades per al procés de *carving* es donaran ordenades pels mínims en la funció àrea, cosa que, com s'ha vist, accelera el procés. Per aquest procés d'esculpit de vòxels, s'ha procurat reduir també el temps d'execució i la mida dels fitxers utilitzant la projecció dels vòxels amb l'ajut del coprocessador gràfic, representant-los en una estructura d'arbre *octree*, acolorint-los i projectant-los damunt un mapa de distàncies com s'ha mostrat en el capítol sisè.

Així doncs, **el mètode d'obtenció de vistes per selecció, reconstrucció 3D per *space carving* i síntesi per interpolació** es planteja definitivament tal com mostra la figura 7.4 (a la següent pàgina). En l'apartat de resultats d'aquesta tesi s'analitzarà la qualitat de les vistes obtingudes, la quantitat de disc necessitada per emmagatzemar la informació i el temps de processador requerit per a la seva execució.

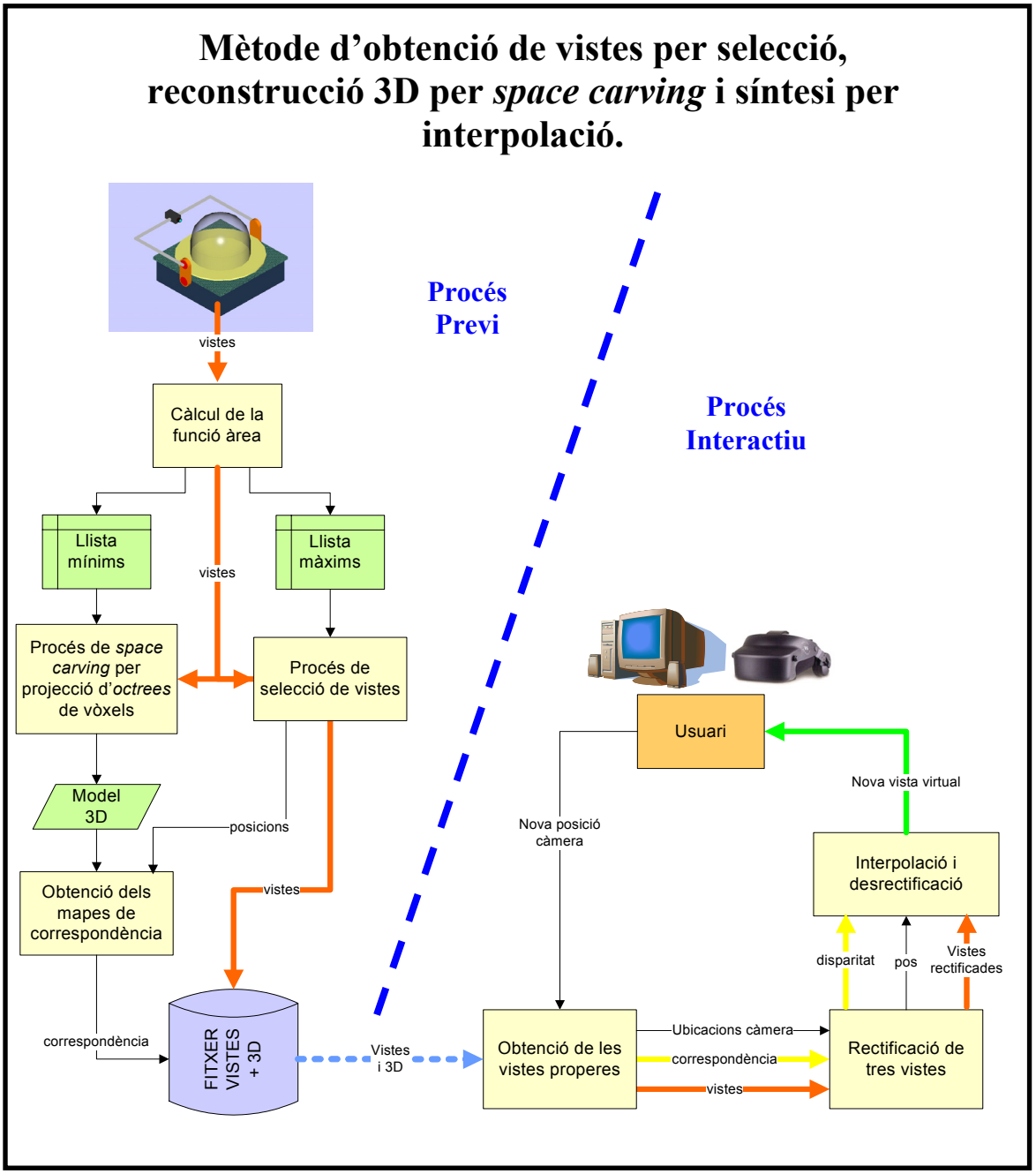


Figura 7.4 El mètode d'obtenció de vistes per aplicacions de realitat augmentada i telepresència a partir de selecció, esculpit del volum tridimensional i síntesi per interpolació de vistes. Consta d'un procés diferit que permet emmagatzemar la informació necessària i d'un procés interactiu on l'usuari podrà utilitzar les noves vistes en la seva aplicació.

8. Anàlisi de la qualitat de les imatges. Fonts d'error

S'han presentat tres mètodes diferents d'obtenció de vistes, un basat en compressió de la informació i accés a les dades, un altre basat en reconstrucció tridimensional, aplicació de textures i dibuix amb el coprocessador gràfic i un tercer basat en selecció d'un conjunt de vistes, càlcul de mapes de correspondència i interpolació de vistes. En els tres casos, el resultat serà una imatge que es presentarà remotament, o damunt un entorn de realitat augmentada. Així doncs, a l'hora d'avaluar la bondat dels resultats de cada un dels mètodes caldrà tenir en compte que el resultat és una imatge i, per tant, cercar criteris de mesura de la qualitat de les imatges. Donat que s'ha generat una gran base de dades de les vistes de l'objecte (veure capítol tercer) aquesta avaluació no serà "a cegues"; es disposa també de la imatge que s'hauria de veure des d'aquell punt de vista i per tant es podrà fer una comparació entre la imatge original i l'obtinguda.

Els autors que han definit mètodes d'avaluació de qualitat d'imatges, ho han fet seguint diferents criteris; per una banda cal distingir l'anàlisi de seqüències d'imatges de l'anàlisi d'imatges estàtiques. Paral·lelament, diferenciarem entre tècniques de mesura d'error orientades a la percepció humana, que tindran en compte aspectes fisiològics i psicològics de les tècniques destinades a mesurar errors per a computació (Taula 8.1).

Criteri de classificació	**Estructura de la informació**		**Destinatari de la informació**	
	Seqüències d'imatges.	Imatges estàtiques	Percepció humana	Màquina de còmput
Propietats	Es té en compte la capacitat de filtrat de l'ull humà. Cerca del "mínim acceptable".	Tècniques més comuns i rigoroses: les mesures dels errors són objectivables.	S'intenta quantificar la subjectivitat dels individus amb la creació d'estadístiques.	Obtindrem mesures d'error numèriques i objectives.

Taula 8.1. Classificació dels mesuradors d'error d'imatges sintètiques.

Quan el destinatari de la informació és un ésser humà, es tendeix a usar un criteri d'apreciació subjectiva, és a dir, es presenta a un grup de persones una o diverses imatges obtingudes segons diferents algoritmes i es creen estadístiques d'apreciació per part dels enquestats [Winkler 99]. D'aquesta manera s'estima la impressió que un producte basat en l'algoritme de síntesi estudiat té en els individus i en conseqüència en un possible mercat. Malgrat la coherència de l'enfocament, la dada no deixa de ser subjectiva i per tant molt difícil d'emprar per a estudis preliminars on, els anàlisis d'error obtinguts per computadors permetran l'automatització dels estudis i la utilització de grans bancs de proves d'imatges.

Serà bo doncs, trobar alguna mesura que permeti comparar numèricament la bondat de vàries imatges o de varis mètodes de síntesi d'imatges, que permetrà fer-ne una classificació correcta i jutjar l'aplicabilitat de cadascun d'ells enfront diversos escenaris.

En l'anàlisi d'error per seqüències d'imatges es tenen en compte aspectes espacials i temporals, així com la capacitat d'integració del cervell humà per elements en moviment. L'establiment de mètriques per la qualitat de vídeo digital [Watson 99] involucra models de processat de l'ull humà [Pappas 99], integració de moviment i filtres en l'espai i el temps. L'aplicació d'aquestes mètriques emprades en la transmissió, codificació i compressió de vídeo, així com els requeriments del mercat, provoca la seva permanent evolució i redefinició (el temps mig d'aparició d'un nou producte per compressió de vídeo digital, com els codificadors MPEG, és de pocs mesos).

Per la comparació d'imatges estàtiques, normalment la imatge sintetitzada amb una de referència, s'utilitzen mesuradors estàndards de comparació de seqüències de nombres i d'altres que tenen en compte el significat d'aquestes dades i comparen també descriptors de les formes, contorns, freqüències existents en la imatge, etc. El treball d'Ahumada per la NASA [Ahumada 93] exposa detalladament les tècniques existents en el moment, que s'han mantingut fins a l'actualitat. L'ús d'aquests mesuradors permetrà obtenir dades objectives de la similitud o dissimilitud de dues imatges. A continuació es parlarà de les fonts d'error tingudes en els diferents processos i es concretaran els mesuradors de qualitat de les imatges que s'empraran.

8.1. Fonts d'error.

Tot i que els tres mètodes d'obtenció de vistes analitzats acaben generant una imatge i es pot fer una comparació de les imatges resultat amb la imatge esperada, s'ha considerat interessant desgranar, per cada un dels mètodes utilitzats, les seves fonts d'error particulars i proposar idees per a reduir-les.

Hi haurà un primer conjunt de fonts d'error, que seran dependents de la plataforma d'adquisició i que afectaran a tots tres mètodes. Aquests **errors comuns** a tots els mètodes proposats seran els següents:

- Errors en l'adquisició i digitalització de la informació fotomètrica. Aquí caldrà comptar amb els errors del captador de la càmera, degut a aberracions de l'òptica, soroll tèrmic al captador i errors d'alineament del sistema òptic, i els errors en la targeta digitalitzadora com la manca de resolució i el soroll electromagnètic. En ambdós casos la utilització de dispositius més cars i de més qualitat reduiria els errors. Els errors d'alineament del captador i la òptica de la

càmera es poden minimitzar mitjançant una adequada calibració dels paràmetres intrínsecs de la mateixa.

- Alineament dels eixos del robot posicionador entre ells, amb l'eix òptic de la càmera i centrat de l'objecte en el sistema posicionador. Tots aquests errors provocarien que, en una seqüència de vistes obtingudes de l'objecte en repòs, aquest adquirís un moviment de balanceig, de rotació o de pivotació respecte a un punt, ja que el sistema suposaria que la càmera real i la càmera virtual segueixen la mateixa trajectòria quan això no és cert. Aquest error constructiu del sistema s'ha procurat minimitzar tal com mostra el capítol tercer, amb la calibració del sistema posicionador, aconseguint trobar les cotes màximes de l'error (veure taula 3.1).

Aquests errors comuns als tres mètodes no podran ser detectats amb la comparació de les imatges ja que també afecten al conjunt d'imatges de referència. A continuació es cercaran les fonts d'error particulars de cada un dels mètodes emprats. En el cas del primer mètode, el que s'ha anomenat de **compressió i accés** a les imatges, la principal font d'error que arribarà a la imatge final serà depenent del mètode de compressió emprat:

- Pèrdua de qualitat en la imatge per l'algoritme de compressió emprat. Els compressors de seqüències d'imatges poden basar-se en mètodes de compressió amb pèrdua de qualitat o sense pèrdua de qualitat. En el primer cas, no s'introduirà cap error al procés degut a la compressió de vídeo, però la mida de les bases de dades a emprar pot ser prohibitiva (veure taula 3.2). En el cas dels algoritmes de compressió amb pèrdua, tots es basen en la detecció de regions mòbils en les imatges, quadriculació de les imatges i transformació freqüencial de les imatges amb eliminació dels coeficients menys significatius. Així doncs, els errors més comuns introduïts són el de filtrat de la imatge i el de pèrdua de resolució espacial i de color.

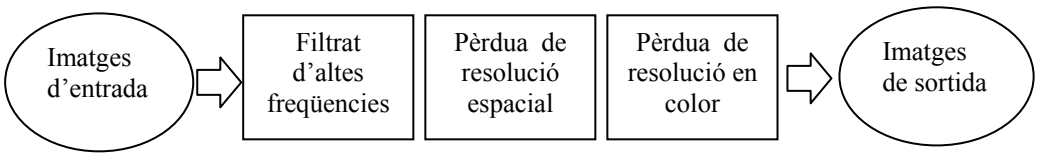


Figura 8.1 Errors típics dels mètodes de compressió d'imatges

En el cas del segon mètode d'obtenció de vistes, consistent en **la reconstrucció tridimensional de l'objecte i l'aplicació de textures**, apareixen fonts d'error particulars que s'enumeren en la següent llista:

- Error degut a la reconstrucció tridimensional de l'objecte. La manca de precisió en l'obtenció del model tridimensional, calculada en el capítol quart, provocarà desplaçaments en la ubicació de punts de la superfície de l'objecte. La manca de resolució en el volum tridimensional emprat provocarà pèrdua de resolució espacial en les imatges sintetitzades i per tant un filtrat en la representació de parts de l'objecte.
- Error degut a la projecció. La discretització del volum ocupat per l'objecte farà que, a l'hora de projectar-lo en una imatge bidimensional apareguin artefactes ja

que el procés seguit pel coprocessador gràfic per combinar o interpolar colors en els píxels mai serà igual al seguit en el món físic.

- Error degut al canvi d'il·luminació i aplicació de textures. Un altre artefacte degut a la projecció d'un model tridimensional és que aquest model es representa informàticament com un conjunt de triangles als quals s'aplica la textura adequada i s'il·lumina segons l'angle format pel vector normal al triangle i la direcció de la font de llum. Aquest conjunt de triangles no tindrà exactament la mateixa inclinació de la superfície de l'objecte el que produirà canvis en la il·luminació de la textura aplicada.

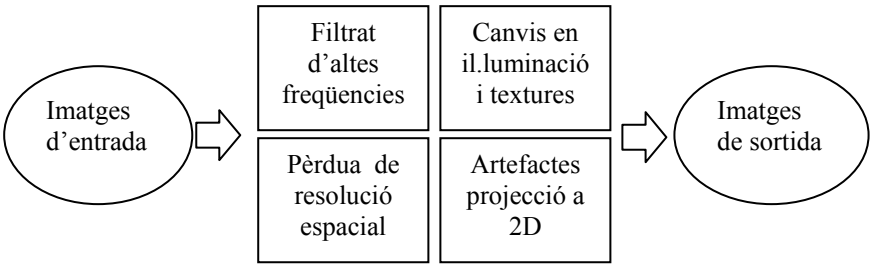


Figura 8.2 Errors típics dels mètodes de reconstrucció tridimensional i aplicació de textures.

Finalment, el mètode d'**obtenció de vistes per selecció i síntesi**, presenta un altre conjunt d'errors específics:

- Error degut a la selecció de vistes. La tria d'un conjunt insuficient o poc representatiu de vistes de l'objecte provocarà que a l'hora de sintetitzar les noves vistes quedin regions senceres buides (fora de la regió estèreo pel parell de vistes font) que calgui omplir per interpolació, provocant pèrdua de resolució en la imatge resultat.
- Error en l'obtenció de la informació tridimensional (correspondència, disparitat) necessària per la síntesi, que provocarà de nou la interpolació de molts punts amb pèrdua de resolució i informació.
- Errors en la computació de les transformacions geomètriques (rectificació i desrectificació d'imatges) necessàries per a la síntesi, que provocarà un filtrat espacial en la imatge amb la pèrdua d'informació d'alta freqüència com són les cantonades dels objectes.
- Modificació de la informació cromàtica. Quan degut als errors mostrats anteriorment dos punts corresponents no tinguin exactament el mateix color caldrà (per exemple) fer-ne el valor mig, modificant la informació de color de la imatge.

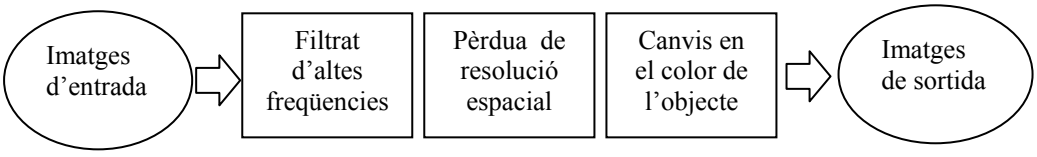


Figura 8.3. Errors típics introduïts pel mètode de síntesi de vistes.

D'aquest conjunt de fonts d'error, es dedueix que les afectacions principals seran la pèrdua d'informació d'altes freqüències (cantonades), la pèrdua d'informació espacial, canvis en el color i aparició d'alguns artefactes en la projecció de models tridimensionals. Per corregir-los caldria doncs, augmentar la resolució de les imatges, augmentar la precisió en la captura o augmentar la resolució volumètrica en els casos en que cal obtenir un model tridimensional. Tota aquesta informació sobre els tipus d'error esperats en els diferents mètodes serà emprada a continuació per seleccionar els millors procediments per comparar i avaluar la qualitat de les imatges obtingudes.

8.2. Anàlisi de la qualitat de les imatges obtingudes.

Un cop determinades les fonts de soroll que afectaran als mètodes d'obtenció de vistes és el moment de triar alguns indicadors de la qualitat de les imatges que permetin comparar-les entre elles. Es farà èmfasi especial en aquells indicadors que estiguin més relacionats amb els efectes de les fonts de soroll identificades. Com que en tots els casos es disposa del que seria la imatge patró (aquella que s'espera tenir per el punt de vista demanat), el més adequat es parlar de **fidelitat de la imatge obtinguda** [Pappas 99]. D'altra banda es triaran les tècniques de comparació de parelles d'imatges estàtiques ja que l'obtenció de vistes no té per que estar lligada als possibles moviments de la càmera. A més a més, els mètodes d'anàlisi d'error per imatges estàtiques són més rigorosos que els emprats en l'anàlisi de seqüències de vídeo.

8.2.1 Mètode subjectiu estadístic

En primer lloc, i donat que el destinatari de les imatges serà sempre un humà, seran preferibles els mètodes subjectius que avaluen el grau de complaença del receptor. Aquests mètodes es basen en l'estudi estadístic de la resposta d'un grup d'enquestats, als que se'ls va presentant el resultat de diferents mètodes. Aquest és el cas dels estudis emprats pels creadors dels estàndards de compressió de vídeo. En la seva darrera versió, l'estàndard mpeg-7, els tests s'han efectuat amb centenars d'usuaris responent quantitativament (1-5) al grau de satisfacció donat per les imatges presentades. En aquesta tesi s'ha emprat també aquest mètode de l'enquesta, seguint aquest patró de puntuació d'u a cinc. Així doncs, aquest serà el primer valor adjuntat als resultats de cada mètode presentat en el capítol següent: el de l'apreciació subjectiva segons el barem presentat a la taula 8.2.

Qualitat/Fidelitat excel·lent	1
Qualitat/Fidelitat bona	2
Qualitat/Fidelitat acceptable	3
Qualitat/Fidelitat dolenta	4
Qualitat/Fidelitat inacceptable	5

Taula 8.2. Mesura emprada per l'anàlisi subjectiu de la qualitat de les vistes.

Del conjunt de respostes obtingut segons aquest barem per cada un dels mètodes se'n farà la mitjana i s'emprarà com a valor de referència de qualitat.

8.2.2 Mètodes numèrics

Deixant de banda aquest primer avaluador, s'havien de triar altres d'objectius on el resultat de la seva aplicació fos una funció de la comparació de dues imatges, fàcil de programar en un sistema informàtic. Els avaluadors més comuns per a la comparació entre parelles d'imatges són els següents:

- Càlcul de la diferència d'imatges, avaluada segons SSD (*Sum of Squared Differences*) o SAVD (*Sum of Absolute Value Differences*) consistent en, per tots els píxels de les dues imatges, computar la seva diferència i fer-ne la suma al quadrat o en valor absolut. També es pot usar algun altre mètode com el càlcul de l'entropia de la imatge diferència. Aquests mètodes permeten trobar variacions en els colors de les imatges i translacions dels seus elements.

- Comparació de PSNR (*Peak Signal-to-Noise Ratio*) que és la relació entre el senyal i el soroll trobat en cada una de les imatges. Quan major sigui aquest valor, més qualitat té la imatge obtinguda.

- Comparació d'histogrames. Malgrat que és un mètode en general poc fiable, ja que és fàcil demostrar que imatges diferents poden tenir el mateix histograma, quan se sap que un grup d'imatges representen la mateixa escena, pot servir per detectar variacions en la il·luminació i certs efectes de filtrat. La utilització de l'histograma conjunt de dues imatges com a mesura de desalineament entre imatges [Hill 94], permet obtenir un valor de dissimilitud calculant el moment de tercer ordre de l'histograma conjunt o la seva entropia.

- Comparació de transformacions freqüencials. Aquests mètodes consisteixen en aplicar alguna transformació matemàtica a la imatge que ofereixi informació de les seves components freqüencials. S'usen principalment tres transformacions:
  - Transformada discreta de Fourier
  - Transformada discreta del cosinus
  - Transformada *wavelet*

  La transformada de Fourier és la més coneguda en matemàtiques i física ja que ofereix la descomposició en una suma infinita de senyals periòdics bàsics de qualsevol senyal d'entrada (en aquest cas, una imatge). La transformada discreta del cosinus i la transformada *wavelet*, que també fan la descomposició d'una imatge en la suma d'infinits senyals bàsics, són emprades actualment en els compressors d'imatge ja que la forma en que surten els seus coeficients és més fàcil d'utilitzar que en el cas dels coeficients de Fourier. En els tres casos, es pot triar un subconjunt finit dels coeficients que permeten reconstruir la imatge amb qualitat suficient. La comparació dels coeficients de dues imatges donarà informació de les freqüències bàsiques que conté cadascuna.

Selecció dels avaluadors de qualitat:

Les figures 8.1, 8.2 i 8.3 mostren que les diverses fonts localitzades en els mètodes d'obtenció de vistes indueixen principalment tres tipus d'error:

- Modificació en els colors de la imatge.
- Pèrdua de resolució espacial.
- Filtrat d'altes freqüències.

En conseqüència, d'entre els comparadors d'imatge presentats i altres (es pot veure el report fet a [Delso 03]) s'han triat aquells que permetran ser més sensibles a aquestes fonts d'error. Els tres avaluadors triats, i que seran utilitzats en el capítol de resultats en la comparació dels diferents mètodes d'obtenció de vistes seran (respectivament al tipus d'error al que són més sensibles):

- Anàlisi de l'histograma conjunt de dues imatges
- Diferència d'imatges
- Comparació de coeficients de la transformada de Fourier

8.2.2.1 Anàlisi de l'histograma conjunt de dues imatges

L'histograma d'una imatge dóna informació del nombre d'aparicions de cada color en la imatge. Si la imatge és en RGB, es generaran tres histogrames un per cada banda de color. De la comparació dels histogrames de la imatge obtinguda per qualsevol dels mètodes amb la imatge original, es podrà deduir una variació en la il·luminació de l'objecte.

L'histograma conjunt de dues imatges, dóna la relació entre els colors dels píxels de les dues imatges, píxel a píxel, creant una matriu bidimensional de [1..MaxColor][1..MaxColor] on cada element (i, j) compta el nombre de vegades que un píxel que a la imatge 1 té el color i, a la imatge 2 té el color j. Matemàticament es defineix així per dues imatges X,Y amb N píxels cada una:

$$HC(i,j) = \sum_{n=1}^{N} \delta_{i,X(n)} \cdot \delta_{j,Y(n)} \quad ; \quad \delta_{i,X(n)} = \begin{cases} 1 & si\ i = X(n) \\ 0 & si\ i \neq X(n) \end{cases}$$

$$\delta_{j,Y(n)} = \begin{cases} 1 & si\ j = Y(n) \\ 0 & si\ j \neq Y(n) \end{cases} \qquad \text{(Eq. 8.1)}$$

Per dues imatges iguals, el resultat de l'histograma conjunt és l'histograma de qualsevol de les dues imatges damunt la recta x = y. Si hi ha hagut modificacions en la il·luminació, aquesta recta es desplaça amunt o avall. Si s'ha produït un filtrat s'eixampla la dispersió dels valors, si hi ha hagut un desplaçament en la imatge també augmenta la dispersió dels valors. Així doncs, l'anàlisi d'aquest histograma conjunt donarà prou informació per detectar alguns dels errors típics ens els processos plantejats. La figura 8.4 mostra un exemple de l'aspecte de l'histograma conjunt per una imatge i ella mateixa després de patir un procés de desenfocament.

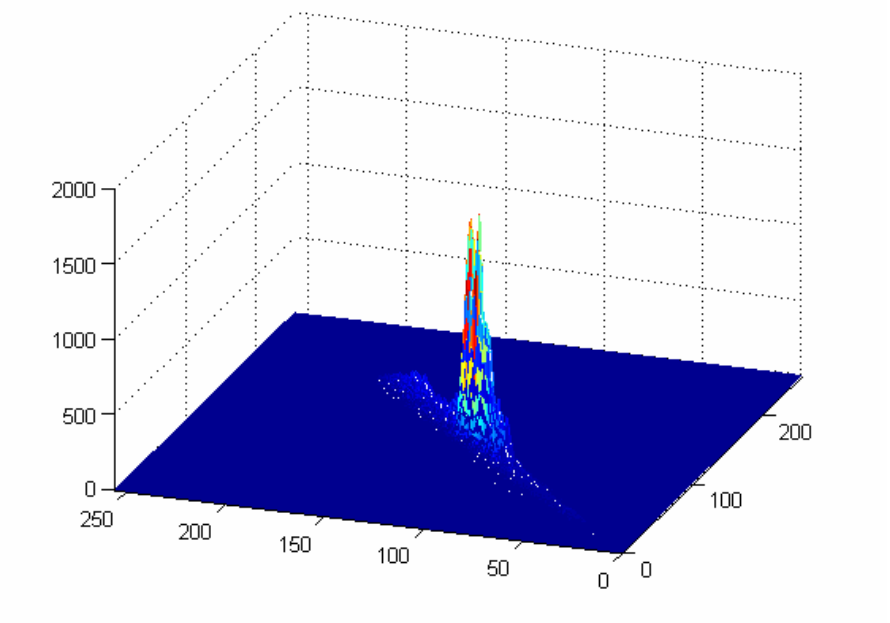


Figura 8.4 Histograma conjunt d'una imatge i ella mateixa després d'un procés de *blur*.

Com a mesura de la dispersió en aquest histograma conjunt [Collignon 95] [Studholme 95] es proposa l'ús de l'entropia, concretament de l'entropia de Shannon [Shannon 48] que matemàticament es defineix com:

$$H = -\sum_{x} p(x) \cdot \log_2 [p(x)] \qquad \text{(Eq. 8.2)}$$

Finalment, i donat que les imatges són tractades en color, es planteja emprar aquesta mesura per cada una de les bandes R,G i B. Per tant, donades dues imatges, una la imatge esperada per un punt de vista I_{OR} i l'altra obtinguda per algun mètode de síntesi, projecció o accés a imatges pregravades comprimides I_{OB} es crearan tres histogrames conjunts entre les dues imatges $HC_R(I_{OR},I_{OB})$, $HC_G(I_{OR},I_{OB})$, $HC_B(I_{OR},I_{OB})$ i per cada un d'ells s'obtindrà una entropia H_R, H_G i H_B que es sumaran per obtenir una mesura de similitud entre les dues imatges.

8.2.2.2 Diferència d'imatges

Donat que la resta d'imatges és la operació menys invariant a la translació, s'ha decidit usar-la per detectar si s'han produït desplaçaments en els processos d'obtenció de vistes. Com a mesura de la diferència entre les imatges, s'ha triat la més senzilla que és la suma de diferències quadràtiques que dóna un valor directament proporcional a la diferència.

$$SSD = \sum_{(i,j)} (I_{OR} - I_{OB})^2 \qquad \text{(Eq. 8.3)}$$

Donat que es tenen tres bandes de color i la SSD està definida per una banda, es farà com en el cas anterior i es sumarà el valor de les diferències en vermell SSD_R, verd SSD_G i blau SSD_B.

8.2.2.3 Comparació de coeficients de la transformada de Fourier

El tercer camí triat per obtenir un valor de la similitud entre les dues imatges és el de l'anàlisi freqüencial via la transformada de Fourier concretament, la transformada discreta ràpida de Fourier bidimensional (FFT2 de l'anglès *Fast Fourier Transform*).

La transformada de Fourier de cada una de les imatges torna una matriu de coeficients on l'element (0,0) representa la component contínua del senyal, els coeficients més propers a aquest representen les baixes freqüències i els més llunyans les altes freqüències (veure figura 8.5). Cal fer notar que la informació apareix replicada en quatre quadrants i tal com mostra la figura, es pot emprar només la d'un d'ells.

Si la imatge generada per algun dels mètodes d'obtenció de vistes ha sofert un procés de filtrat en altes o baixes freqüències, o en l'extrem, un augment de la component contínua (que representaria un augment de la lluminositat mitjana), es pot detectar per la comparació dels coeficients de Fourier. Per tant, el que es farà és, per les dues imatges I_{OR} i I_{OB}, calcular les seves respectives transformades de Fourier amb la FFT2, F_{OR} i F_{OB} i obtenir tres valors de la diferència: un per la component contínua, un pel promig de les components de baixa freqüència i un pel promig de les components d'alta freqüència. La transformada de Fourier sols s'aplicarà damunt la component lluminositat de les dues imatges.

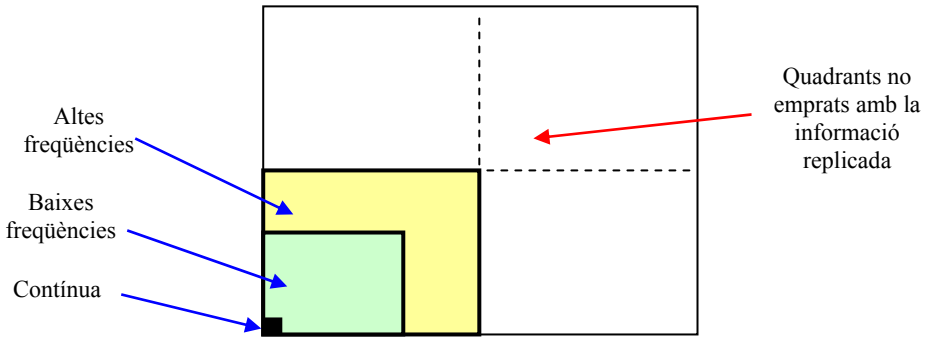


Figura 8.5 Definició de les àrees d'interès en el resultat de la transformada de Fourier.

Per acabar aquest apartat, a la figura 8.6 es mostra l'aplicació de la FFT2 damunt una imatge de la urna funerària romana i la mateixa imatge en la que s'ha reduït la lluminositat. El requadre inferior de la figura presenta la diferència de valors de les components contínua, de la mitjana de components de baixa freqüència i de la mitjana de components d'alta freqüència. El valor numèric és fàcil d'interpretar en el cas de la diferència de components contínues (increment o decrement de la il·luminació mitjana). En el cas de les freqüencials, s'emprarà només com a eina de comparació.

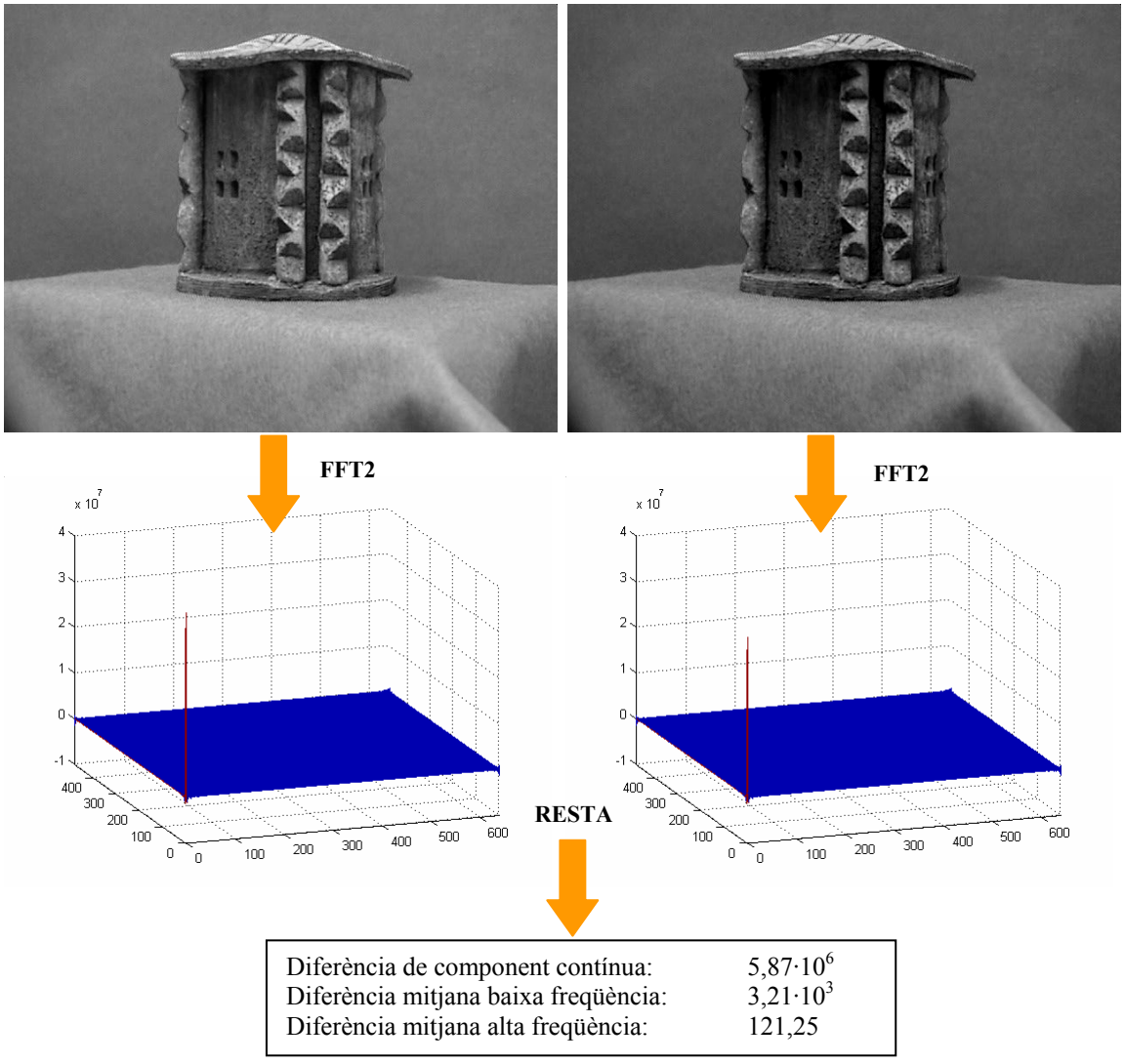


Figura 8.6 Exemple d'utilització de la transformada discreta de Fourier per avaluar diferències entre dues imatges.

9. Comparació de mètodes, resultats i aplicacions

En aquest capítol es mostraran els resultats dels diversos experiments realitzats durant l'elaboració d'aquesta tesi; els experiments s'han emprat a vegades per corroborar o refusar conjectures i altres vegades dels propis experiments n'han aparegut de noves. S'ha intentat abordar els experiments sense prejudicis, per no caure en errors d'apreciació "es un error teoritzar abans de tenir dades, es corre el risc d'emmotllar les dades a la teoria, i no les teories a les dades" [Conan-Doyle 1891], especialment en la comparació dels diferents mètodes d'obtenció de vistes. A continuació es presenten aquests resultats juntament amb una avaluació dels seus requeriments temporals a data de primavera de 2006, com sempre, l'evolució tecnològica farà que aquests temps es redueixin dràsticament en els propers anys.

9.1 Obtenció de vistes d'objectes reals.

Els tres mètodes presentats d'obtenció de vistes d'objectes reals, tenen naturaleses diferents, per això la seva comparació ofereix dificultats ja que cadascun d'ells tendeix a utilitzar recursos diferents: disc, coprocessador gràfic (GPU) o processador principal (CPU). A continuació es mostren tots tres i una comparativa dels seus rendiments. La interpretació del seu significat es deixarà pel capítol de conclusions.

9.1.1 Pel mètode d'accés a fitxers de vídeo.

L'element clau en aquest mètode d'obtenció de vistes és la utilització de fitxers de pel·lícula de vídeo emmagatzemats en disc. Com s'ha vist, si es disposa de totes les vistes de l'objecte sols farà falta indexar aquest fitxer per tenir la vista requerida. El mètode és, evidentment, el més senzill però té una necessitat ingent de capacitat de memòria. Això fa que, actualment només els medis d'emmagatzemament a disc puguin suportar aquestes capacitats. Això, juntament amb el fet de que la pel·lícula sigui enregistrada com una seqüència implica que els temps d'accés no seran homogenis. L'accés des de la imatge *n* a la imatge *n+1* o *n-1* (la dreta o l'esquerra) serà més ràpid que l'accés a la imatge *n+vpv* (de vistes per volta) o *n-vpv* on hi ha la vista superior o la inferior. El temps d'accés serà doncs un paràmetre a tenir en compte juntament amb la mida dels fitxers a disc i la seva qualitat.

La figura 9.1 mostra la mateixa vista d'un objecte, amb diferents compressions aplicades, les quals implicaran diferents mides a disc i qualitats. Malgrat que la relació

entre la qualitat de la vista i la mida a disc és dependent de l'objecte i de la tècnica de compressió emprada, s'ha observat un decaïment similar de la qualitat en funció del factor de compressió. També s'ha constatat que les variacions d'aquesta relació dependents de l'objecte són mínimes pels experiments realitzats.

Figura 9.1 Aspecte de la vista obtinguda amb un compressor de vídeo DivX amb compressió de 75% (esquerra), 85% (centre), 95% (dreta)

Segons el factor de compressió mostrat (s'ha triat el DivX per ser el de millors prestacions en el moment de realització de la tesi) s'obtenen uns fitxers de major o menor mida (veure figura 9.2). En alguns casos, el fitxer només pot ser gravat en discs durs o cintes magnètiques; a partir de certs factors de compressió es pot usar també emmagatzemament òptic en discs de 8GB o 4,7GB (formats DVD) o 800MB (formats CD). Això permetria tenir una base de dades d'un disc per objecte però, el temps d'accés a les vistes no seqüencials en medis òptics impossibilita de moment la seva utilització en aplicacions interactives.

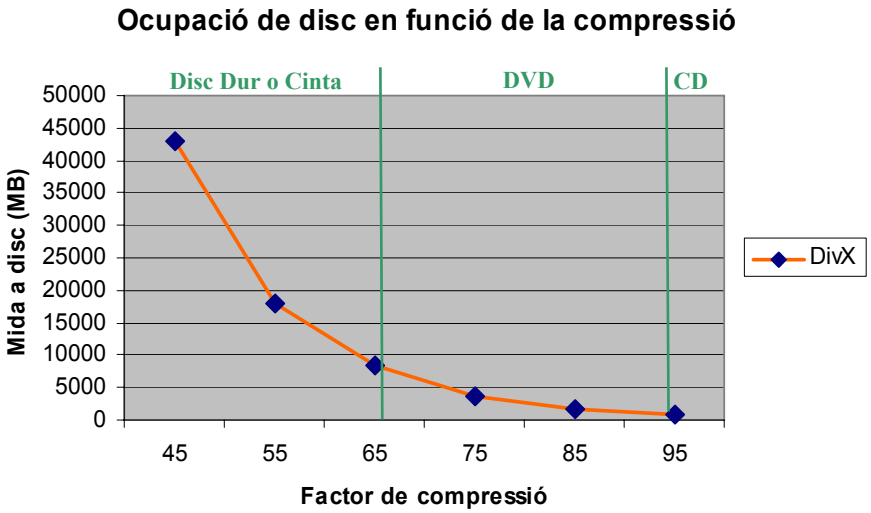


Figura 9.2 Mida dels fitxers d'emmagatzemament de 1200x300 vistes en funció del factor de compressió emprat.

Cal fixar-se ara en el temps d'accés a la informació. Aquest temps serà en principi, dependent del medi d'emmagatzemament físic, del temps de descompressió requerit i de la càrrega del sistema.

- El medi físic (memòria, disc dur, CD o DVD) tindrà un temps d'accés que depèn de la seva tecnologia. Si el nombre de vistes és petit, el millor és

emprar la memòria dinàmica del sistema per guardar-hi totes les vistes descomprimides, l'únic problema és que en 1GB de memòria (part raonable a dedicar a les vistes en un computador actual), sols caben unes 1200 vistes, les justes per fer una volta a l'objecte amb la resolució capturada. Si es volen tenir totes les vistes en memòria, o tenir més d'un objecte, caldrà recórrer als discos. La taula 9.1 mostra els temps d'accés promig a una vista per tres sistemes d'emmagatzemament diferents: memòria dinàmica, disc dur i DVD.

- El temps de descompressió de la vista obtinguda del fitxer, sols serà aplicable als casos del disc dur o DVD, en la memòria física, ja es té descomprimida. La taula 9.1 mostra també el temps estimat d'execució de l'algoritme de descompressió.
- Es considerà que la càrrega del sistema no afecta als experiments. En un sistema operatiu que no sigui de temps real (Linux, Windows) no serà predictible i l'únic que es pot fer és alliberar al màxim de tasques no necessàries al sistema.

	Memòria (DDRAM)	Disc magnètic (HD)	Disc òptic (DVD)
Temps d'accés a vista (promig)	<1ms	1-50*ms	1-700*ms
Temps de descompressió	10-300*ms	10-300*ms	10-300*ms

Taula 9.1 Temps d'obtenció de les vistes en funció del medi físic emprat per emmagatzemar-les.
(*Accés a vistes consecutives - accés a vistes no consecutives)

Manca ara fixar-se en la fidelitat a l'original de les imatges obtingudes en funció del factor de compressió emprat. Com a avaluador d'aquesta qualitat es mostren, a la taula 9.2 el resultat del criteri subjectiu, el de l'entropia de l'histograma conjunt, el de la diferència d'imatges SSD i el de comparació de coeficients de la transformada discreta de Fourier (en el capítol vuitè s'ha explicat el significat d'aquests avaluadors).

Factor de compressió aplicat (%)	Mitjana de l'avaluació subjectiva (1-5)	Diferència d'imatges SSD	Entropia de l'histograma conjunt	Comparació coeficients Fourier		
				Part Continua	Baixa Freqüència	Alta Freqüència
15%	1,1	7M	0,9503	5691	3952	5034
35%	2,0	18M	0,8919	12397	7852	8734
55%	2,9	38M	0,6612	37034	18499	21048
75%	3,6	78M	0,3893	65320	34204	47014
95%	5,0	150M	0,1216	123041	51010	95128

Taula 9.2 Avaluació de la qualitat de les vistes comprimides per cinc factors de compressió emprant els criteris: avaluació subjectiva (mitjana de les respostes de 20 individus), diferència entre imatges (emprant l'avaluador SSD), entropia de l'histograma conjunt i comparació de coeficients de la transformada de Fourier (per la component contínua, la mitjana de les de baixa freqüència i la mitjana de les d'alta freqüència).

Queda clar que pel mètode d'accés a vistes, existeixen dos problemes principals: la quantitat d'informació a emmagatzemar i el temps d'accés i descompressió de les vistes quan aquestes impliquen salts en la seqüència. L'evolució tecnològica fa preveure que la quantitat de memòria necessària, tard o d'hora, no serà un problema, tot i que cada cop es demanaran vistes amb més qualitat que, en contrapartida, faran augmentar el volum de dades necessari. La tecnologia també permetrà accedir més ràpidament a les vistes, l'aparició de bussos de connexió de discs durs cada cop més ràpids com els *Ultra Wide SCSI* i *RAID* i el continu increment de la quantitat de DRAM disponible ho fan pensar. De totes maneres, amb la tecnologia actual, es proposen aquestes millores per a fer compatible el procés amb aplicacions interactives:

- Precàlcul de les vistes més probables: es proposa tenir un procés paral·lel al principal de l'aplicació, llegint de disc i deixant en un espai de la memòria dinàmica paquets de vistes veïnes a la visualitzada en el moment, millorant el rendiment mig de l'aplicació.
- Càrrega dels fitxers de vistes a memòria, sense descomprimir: és clar que les vistes descomprimides no es poden tenir actualment a la memòria principal del sistema. El que si que es pot tenir és, en un disc muntat a la memòria DRAM (l'anomenat *ramdrive*) el fitxer de vistes comprimit, estalviant la part costosa d'accés a les vistes a disc (veure taula 9.1).
- Millora de l'organització de les vistes a disc. Donat que els discs durs poden tenir diversos capçals llegint alhora les mateixes pistes i sectors, en diferents discs, una organització del fitxer de vistes en que les vistes consecutives dreta i esquerra estiguin en sectors veïns i les vistes superiors i inferiors en discs veïns, reduiria també el temps d'accés a la informació. Els sistemes operatius més comuns no afavoreixen però aquesta tasca d'organització de les dades a disc.

9.1.2 Per aplicació de textura a models tridimensionals.

El mètode d'obtenció de vistes per aplicació de textures a models tridimensionals basa el seu funcionament en la utilització dels recursos capacitat de processat i memòria, del coprocessador gràfic (GPU) present en els ordinadors personals. La quantitat d'informació que cal gravar per descriure un objecte és realment petita, sobretot en comparació amb el mètode anterior. Per un objecte com els descrits, representats amb *resX* punts horitzontals per *resY* punts verticals, i considerant que els nombres reals es poden representar suficientment en 32 bits, cal gravar:

$$mida\ fitxer = resX \cdot resY \cdot \begin{Bmatrix} 3\ nombres\ reals : x, y, z \\ 3\ bytes\ de\ color : r, g, b \end{Bmatrix} = 15 \cdot resX \cdot resY\ bytes \qquad \text{(Eq. 9.1)}$$

Que per un objecte de 1200 x 400 punts tridimensionals, fa que el fitxer ocupi 7,3 MB a disc. Aquesta informació pot ser comprimida amb qualsevol mètode de compressió sense pèrdua, obtenint una mida de 2,95 MB per l'objecte urna funerària. Un objecte de 1200 x 400 punts tridimensionals defineix una malla de 960.000 triangles que cal càrregar a la memòria del coprocessador gràfic, on ocupa uns 48MB de memòria, cosa que no representa cap problema per la GPU que actualment es sol equipar amb 128MB-256GB de VRAM (memòria de vídeo).

Es té doncs una relació entre la mida del fitxer de dades que emmagatzema l'objecte i el nombre de triangles que el representaran tal com mostra la figura 9.3.

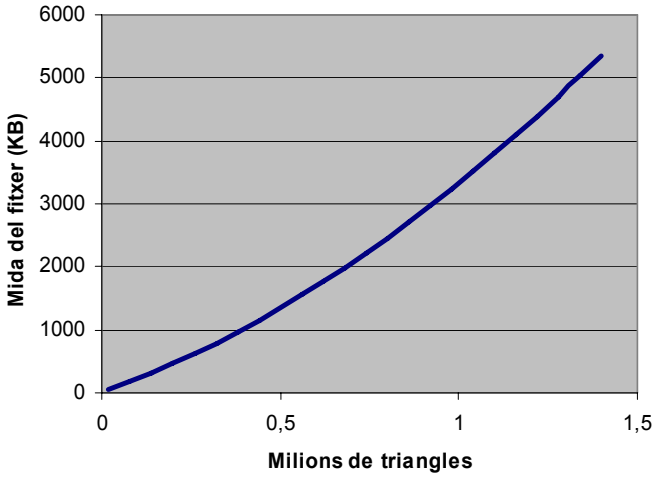


Figura 9.3 Relació entre la mida de les dades emmagatzemades per objecte i el nombre de triangles representats per la projecció del model tridimensional.

La capacitat de representació del coprocessador gràfic, fa que els aproximadament 1.000.000 de triangles de l'objecte siguin projectats a la pantalla en un temps de 200ms, obtenint un refresc de 3Hz. La taula 9.3 mostra la capacitat de projecció màxima, en triangles per segon, de diversos dispositius GPU del mercat [NVidia 06][ATI 06][XBox 06][SG 06]:

Nom del dispositiu	Capacitat de projecció teòrica màxima (triangles/s)	Altres característiques
Ordinador personal, targes convencionals	20-70 M	64MB RAM, 100MHz GPU
Ordinadors personals, targes alt rendiment jocs	100-200 M	128-256 MB RAM, 200MHz GPU
Workstation pensades per tasques de CAD	50-100 M	128-256 MB RAM, 200MHz GPU
Consoles de jocs domèstiques	500 M	512MB RAM, 500MHz GPU

Taula 9.3 Capacitat de representació en triangles per segon de diversos dispositius GPU que es troben al mercat

Manca parlar de la qualitat de la imatge obtinguda mitjançant el procés d'aplicació de textures a models tridimensionals (taula 9.4).

Nombre de triangles (en milers)	Apreciació subjectiva	Error mesurat per SSD	Temps de projecció (en ms)
200	4,0	152M	85
400	3,1	135M	108
600	2,5	113M	139
800	2,1	97M	160
1000	1,8	89M	198

Taula 9.4 Avaluació de la qualitat de la imatge obtinguda pel criteri subjectiu (mitjana de les respostes de 20 individus) i per la diferència SSD i temps de projecció del model tridimensional de l'escena al coprocessador gràfic.

És obvi que, en funció del nombre de triangles que es puguin projectar, la qualitat de la vista obtinguda augmentarà i també el temps necessari per la seva projecció. La figura mostra la relació entre el nombre de triangles, la qualitat obtinguda segons el paràmetre subjectiu i el paràmetre SSD i el temps necessari per la seva representació.

Sols queda ara plantejar-se una qüestió: quan augmenti el nombre de triangles que poden manegar les GPU i la seva velocitat de projecció, arribarà a ser indistingible el resultat d'un procés de *render* i una fotografia per a un observador humà? La resposta comunament acceptada és que si, així doncs, és possible que en aquest mètode, quan la informació d'un píxel correspongui aproximadament a la projecció de deu triangles (representaria tenir l'objecte amb una resolució 50 vegades major), s'haurà assolit una fidelitat absoluta respecte a la vista original. Per això caldrà:

- Esperar l'augment de la capacitat gràfica de les GPU (pocs anys).
- Ser capaç de generar un model suficientment precís de l'objecte, augmentant la resolució de la càmera, la precisió del sistema d'adquisició i el nombre de vòxels que es tractin en el procés d'esculpit.

9.1.3 Per selecció i síntesi de vistes.

El mètode de selecció i síntesi de vistes pot oferir qualsevol vista d'un objecte a partir d'un conjunt reduït de vistes i els seus mapes de disparitat. Així com el punt crític dels mètodes anteriors són la quantitat d'informació necessària i el temps d'accés a disc i el nombre de triangles a representar, en aquest cas, el punt crític és l'ús extensiu del processador principal del computador. En la fase prèvia d'adquisició, selecció de vistes, i reconstrucció de la informació tridimensional pel mètode de *space carving* caldrà emprar d'altres recursos del sistema:

- El procés de selecció de vistes requerirà processar totes les vistes disponibles de l'objecte emmagatzemades a disc i fer-ne la segmentació i determinació de la característica àrea.
- El procés de *space carving* requerirà emprar la GPU per la projecció del model de vòxels
- En el procés de *space carving* caldrà definir una estructura de memòria pels vòxels de fins a 1GB (en el cas d'un cub de 1024 vòxels de costat).
- El còmput dels mapes de correspondència a partir del model tridimensional requerirà empra també la GPU per la projecció del volum a les diferents vistes seleccionades.

El resultat de tots aquests passos és un fitxer amb el conjunt de vistes seleccionades i els mapes de correspondència necessaris per la síntesi. La mida d'aquest fitxer, per un nombre de vistes seleccionat *#vistes* i unes imatges de *imX* per *imY* píxels amb 24 bits de color per píxel serà de:

$$mida\ \ fitxer = \#vistes \cdot imX \cdot imY \cdot (24 + 2 \cdot L_2(imX) + 2 \cdot L_2(imY)\ )\ bits \qquad \text{(Eq. 9.2)}$$

Fórmula que surt de comptar que cada píxel de cada imatge haurà de tenir una correspondència amb dues imatges per poder fer la interpolació en el sentit horitzontal i vertical. Per representar una correspondència caldrà el logaritme en base 2 bits del nombre a referenciar. Aquesta informació és, naturalment, susceptible de ser sotmesa a compressió; per l'exemple de la urna funerària romana, s'han comprimit les imatges

amb un algoritme de compressió *jpeg* amb pèrdua i les dades de correspondències amb un algoritme de compressió sense pèrdua, obtenint, per cada vista: 45 KB + 62 KB d'informació, que per les 60 vistes seleccionades de l'objecte, fa un total de 6,3 MB de dades a emmagatzemar pel mètode de selecció i síntesi de vistes.

La taula 9.5 mostra el temps d'execució del conjunt de processos previs a la síntesi de vistes. El fet de que no sigui un procés interactiu fa que es pugui deixar executant sense restriccions temporals. Donat que el temps d'adquisició de les vistes amb el sistema robotitzat és de l'ordre de quatre hores, s'aprofita aquest temps per anar realitzant la selecció de vistes i reconstrucció tridimensional.

Subprocessos	Temps d'execució en PC
Processat de les vistes adquirides (segmentació + àrea)	(1200·300)·20ms = 2 hores
Selecció del conjunt òptim de vistes	2,4 s
Obtenció del model 3D per *space carving*	12,2 s
Generació dels mapes de correspondència	6 s
Gravació de la informació a disc	1 s

Taula 9.5 Temps d'execució dels processos previs per la selecció de vistes i creació dels mapes de correspondència.

Per la part del procés interactiu, caldrà llegir aquesta informació, portar-la a memòria del processador i interpolar les noves vistes. Les dades d'imatges i correspondències, un cop ubicades a memòria, ocupen 128 MB, que són perfectament suportables per un ordinador actual.

Cada cop que arriba una petició de nova vista, la part interactiva de l'aplicació haurà de generar el pla de reprojecció i les matrius associades d'homografia (veure capítol quart), fer la projecció de les vistes al pla (*warping*), sintetitzar la nova vista i aplicar-hi una interpolació per omplir els punts dels que no s'ha pogut obtenir informació. La taula 9.6 mostra el temps d'execució d'aquest procés amb una aplicació de test optimitzada en rendiment damunt de dues plataformes de còmput diferents.

Subprocessos	PC P-IV 1,8GHz 512MB	PC P-IV 2x3GHz HT 2GB
Crear pla de reprojecció i matrius d'homografia	0,0489 ms	0,0192 ms
Projecció de les imatges originals al pla (*warping*)	54,121 ms	31,237 ms
Síntesi de la nova vista (pas de correspondència a disparitat i interpolació)	342,12 ms	152,65 ms
Emplenat dels forats en la nova vista per interpolació	4,0232 ms	3,0102 ms
Temps total procés interactiu	400,313 ms	186,916 ms

Taula 9.6 Temps d'execució del procés interactiu de síntesi de vistes en dues plataformes PC: un *intel Pentium* IV a 1,8GHz amb 512 MB de DDRAM i un intel Pentium amb dos processadors a 3GHz amb tecnologia *hiperthreading* i 2 GB de DDRAM. Mesures obtingudes amb el HPT (*High Performance Timer*).

Es dedueix de la taula mostrada que el procés de síntesi de vistes pot oferir, amb la tecnologia actual un rendiment de 5 o 6 imatges per segon. Possibles optimitzacions en el codi i l'evolució tecnològica fan pensar que en pocs anys s'assolirà la velocitat de 25 imatges per segon, compatibles amb l'anomenat *video-rate*, que defineix allò que l'ull humà percep com a seqüència contínua d'imatges.

La qualitat de les imatges obtingudes estarà lligada amb la bondat dels mapes de disparitat emprats en el procés de síntesi i en conseqüència de la qualitat de la recuperació de la informació tridimensional de l'objecte. De fet, si el mapa de disparitat és dens i no hi ha oclusions, forats o alteracions en la ordenació esquerra-dreta dels píxels, el mètode de síntesi de vistes garanteix que la vista obtinguda serà totalment fidel a l'original. La figura 9.4 mostra dos dels mapes de disparitat obtinguts per l'objecte urna funerària gràcies a la utilització de la reconstrucció tridimensional feta prèviament. Per obtenir unes vistes sintètiques encara més fidels a l'original caldria aplicar les següents millores:

- Fer més precís el mapa de disparitat, augmentant la resolució i precisió del mapa de vòxels obtingut en el procés previ.
- Resoldre els problemes creats pels forats i les oclusions en la conversió del model tridimensional a mapes de correspondència.
- Fer òptim el procés d'interpolació bidimensional un cop obtinguda la nova vista. En l'exemple de les figures 9.5 i 9.6 s'ha aplicat una simple interpolació lineal.

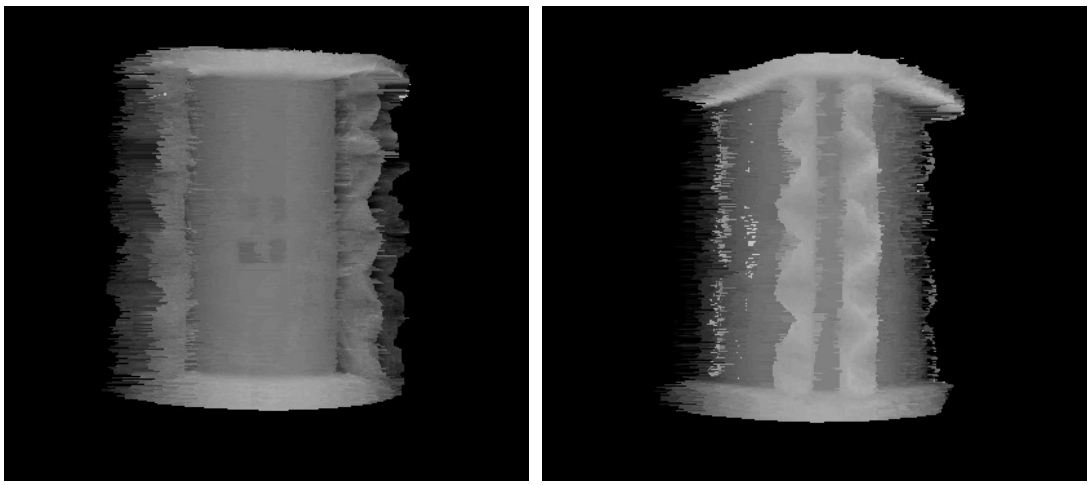


Figura 9.4 Mapes de disparitat obtinguts per interpolar vistes a 0º i 60º respectivament. L'ús del model tridimensional previ ha permès obtenir un mapa de disparitat dens.

La figura 9.5 mostra el resultat del procés de síntesi sense interpolar, mentre que la figura 9.6 mostra el resultat després d'aplicar una interpolació lineal entre els píxels. En la vista "en brut" s'han mostrat en vermell els punts on el procés de síntesi de vistes no ha pogut generar informació degut a oclusions en el moment de la captura, forats, etc. Apareix també una estructura geomètrica degut al procés de projecció al pla. En la vista interpolada es mostra la vista tal com es presentaria a l'usuari, després d'interpolar els punts on mancava informació i fer la desrectificació i el reescalat de la vista per fer-la correspondre a la posició i orientació de la càmera virtual.

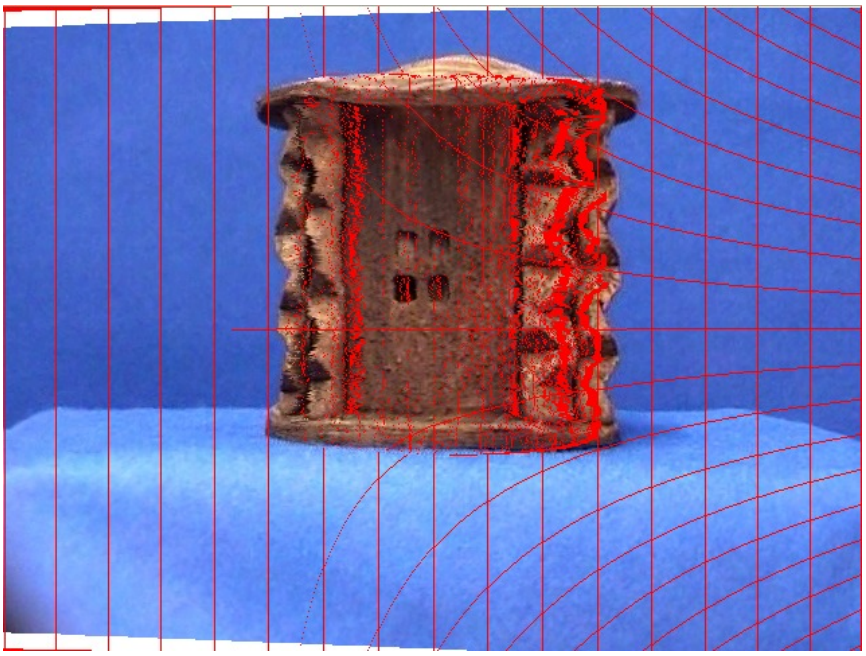


Figura 9.5 Vista obtinguda durant el procés de síntesi. La trama vermella regular representa els punts dels que no es té informació degut a la reprojecció de les imatges o *warping*, els punts vermells dins de l'objecte mostren punts on el mapa de correspondència ha fallat degut a oclusions.

Figura 9.6 Vista interpolada, amb l'escalat final i sense la rectificació del procés de síntesi, tal com es mostraria a l'usuari.

La qualitat de les vistes sintètiques té una dependència del nombre de vistes seleccionades i l'estructura de l'objecte. Com s'ha vist en el capítol sisè, objectes d'estructura senzilla, com per exemple un cub, podran generar correctament qualsevol vista sintetitzada a partir de només vuit vistes originals enregistrades. Pel cas de la urna funerària mostrada s'han gravat 60 vistes amb els seus mapes de correspondència.

9.1.4 Comparació de mètodes.

Els mètodes presentats d'obtenció de vistes acaben generant, per diferents camins, la vista necessària per un sistema de realitat augmentada o telepresència. La primera comparació que es pot fer entre els mètodes es visual; la figura 9.7 mostra la imatge original per un punt de vista, i les obtingudes pels tres mètodes estudiats. En els quatre casos és la imatge corresponent a 9º de longitud i 0º de latitud segons s'han definit en el capítol tercer.

(a) Imatge de referència de l'objecte obtinguda amb el sistema d'adquisició.

(b) Imatge obtinguda per accés a vistes comprimides en un fitxer de vídeo

(c) Imatge obtinguda per interpolació de vistes, amb vistes originals seleccionades a 0º i 18º, la sintètica ha estat generada a 9º.

(d) Imatge obtinguda per projecció d'un model tridimensional de 1200x400 punts, que genera un milió de triangles.

Figura 9.7 Vista original per una posició donada de 9º,0º (a) i resultat dels tres mètodes d'obtenció de vistes proposats: accés a vistes gravades a disc (b), selecció i síntesi de vistes (c) i projecció de models tridimensionals amb textura (d).

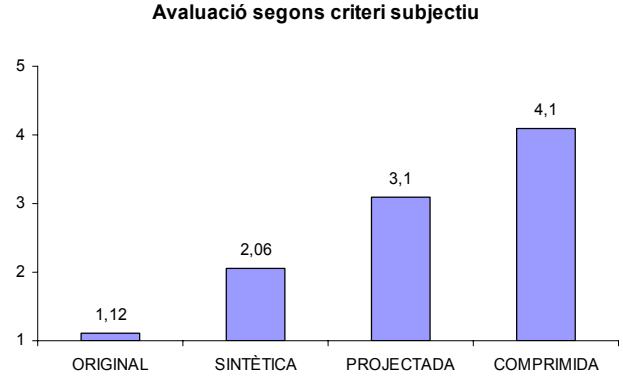
La segona comparativa que es presenta és, pels tres mètodes, una taula amb els temps d'execució dels processos previs i el procés interactiu. La part relativa als temps d'execució dels processos previs es mostra només a tall informatiu, mentre que la part relativa als processos interactius és la que permetrà calcular el nombre d'imatges per segon que proporcionarà cada sistema.

	Mètodes d'obtenció de vistes		
Parts de cada mètode	**Accés a vistes guardades en fitxers de vídeo.**	**Projecció de models tridimensionals**	**Selecció i síntesi de vistes per interpolació**
Adquisició de les vistes de l'objecte	4h	4h	4h
Compressió de les vistes en fitxer	4h	0	0
Còmput funció àrea i selecció de vistes	0	2h	2h
Reconstrucció de l'estructura 3D	0	12-20s	12-20 s
Creació dels mapes de correspondència	0	0	6s
Gravació informació de vistes	0	0	1s
Temps total pels processos previs	**8h**	**6h**	**6h**
Accés a les dades emmagatzemades	30ms	100ms[1]	100ms[1]
Descompressió de la informació	300ms	800ms[1]	250ms[1]
Síntesi de la nova vista de l'objecte	0	0	186ms
Projecció del model tridimensional	0	200ms	0
Dibuix de la nova vista obtinguda	5 ms	0	5ms
Temps total procés interactiu (per vista)	**335 ms**	**200ms**	**191ms**
Freqüència màxima obtinguda (vistes/segon)	3Hz	5Hz	5Hz

Taula 9.7 Taula comparativa dels temps d'execució dels processos previs i interactius pels tres mètodes presentats. Finalment es mostra la freqüència màxima obtinguda d'obtenció de vistes per cada un dels mètodes.

S'observa que cap dels tres mètodes aconsegueix, amb la tecnologia actual, arribar a sintetitzar vistes a *video rate* (25 vistes per segon) amb la qualitat i el nombre de vistes total requerits. A l'hora de comparar la fidelitat de les imatges obtingudes pels diferents mètodes, respecte la imatge original, s'ha intentat equilibrar la quantitat d'espai a disc que es permet ocupar a les dades de cada mètode. La comparativa que es mostra als gràfics de la figura 9.8, que correspon a l'avaluació d'una seqüència d'imatges com les mostrades a la figura 9.7, ha estat per un conjunt de dades d'entrada pel mètode d'accés a vistes de 17 MB (fitxer AVI), pel mètode de projecció de models amb un conjunt de dades de 2,9 MB (fitxer comprimit de punts *x, y, z* i colors) i pel de selecció i síntesi de vistes de 6,3 MB (fitxer de vistes i disparitat). Tot i diferir

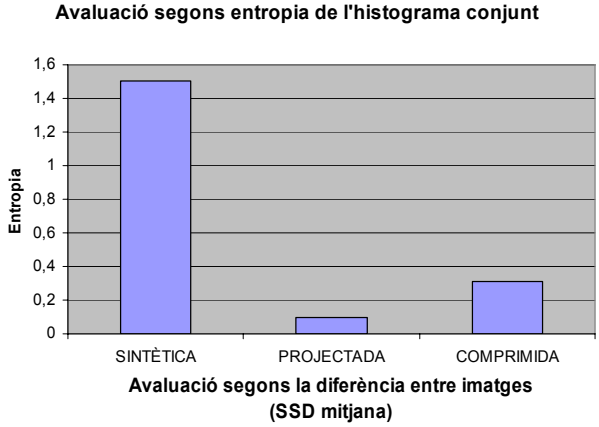
lleugerament les xifres, s'ha intentat dur les mides cap al mateix ordre de magnitud: la d'un fitxer estàndard portable entre ordinadors.



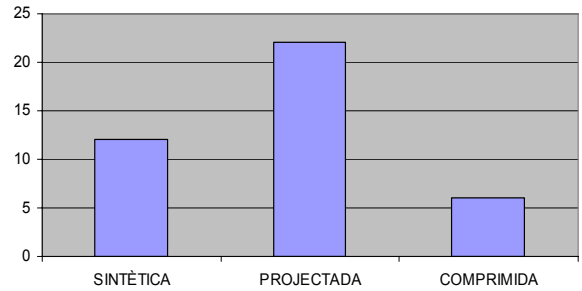
(a) Avaluació segons el criteri subjectiu per una mostra de vint persones. Les valoracions podien ser:

1. Excel·lent
2. Bona
3. Acceptable
4. Dolenta
5. Inacceptable

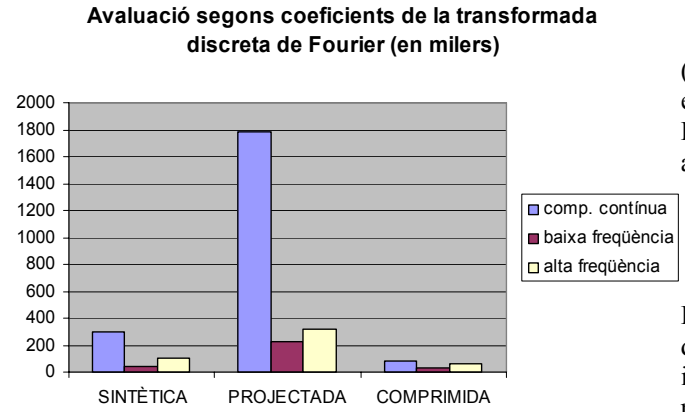
Es mostra la valoració mitjana de la imatge original, presa com a control i de les obtingudes pels tres mètodes.



(b) Comparació de les imatges resultants segons l'entropia de l'histograma conjunt. Valors majors indiquen major fidelitat de la imatge a l'original.



(c) Comparació de les imatges resultants segons la diferència entre imatges. S'ha calculat la suma de diferències al quadrat de les parelles d'imatges (original, mètode) i s'ha dividit pel nombre de píxels. Així, un valor 11 implica la distància mitjana del píxel respecte al seu valor original (en l'espai RGB).



(d) Comparació de les imatges segons els coeficients de la transformada de Fourier. Valors majors indiquen més alteració respecte la imatge original.

Exemple: Un valor de 200.000 en la diferència de components contínues indica que cadascun dels 400.000 píxels ha variat la seva lluminositat en promig 0,5.

Figura 9.8 Comparació de la fidelitat a la original de les vistes obtingudes pels diferents mètodes segons els avaluadors: apreciació subjectiva per la mitjana de 20 respostes (a), entropia de l'histograma conjunt (b), diferència d'imatges per SSD (c) i comparació de components freqüencials amb la transformada discreta de Fourier (d).

9.2 Aplicacions en realitat augmentada.

L'objectiu dels mètodes d'obtenció de vistes presentats, és i ha estat, el d'obtenir d'una manera eficient vistes d'objectes reals per a poder-les introduir en entorns de realitat augmentada. Per això caldrà prendre la vista, si cal, retallar-la, i superposar-la en el lloc precís de l'escena real observada. El càlcul d'aquest lloc precís d'inserció implica un coneixement de l'estructura del món real observat (en el procés interactiu) i el seu alineament amb el món virtual que conté l'objecte del que es presentaran les vistes. Això és el que s'anomena posada en correspondència (*registration* en anglès). A continuació es mostraran els experiments realitzats en un senzill entorn de realitat augmentada, amb el mètode de posada en correspondència triat. Després es mostrarà una descripció sencera d'un **sistema de realitat augmentada amb vistes d'objectes reals**, en funció del mètode de síntesi triat i es farà una avaluació del seu rendiment.

9.2.1 Vistes d'objectes reals en correspondència.

Pels experiments en realitat augmentada, s'ha implementat una plataforma on una càmera determina la seva posició a l'espai a partir de marques conegudes, crea un nou sistema de coordenades referenciat a les marques i hi col·loca els objectes virtuals. Per la seva realització ha calgut afrontar el problema de localització, anomenat LDP (de *location determination problem* [Fischler 81]) i concretament s'ha triat la solució del problema per tres punts, conegut com P3P (de *Perspective-three-point problem*). En aquest plantejament, a partir de tres punts de l'espai de posició coneguda A, B i C, que es localitzen a la imatge com A', B' i C', es resol la posició de la càmera (punt observador). La figura 9.9 mostra com es representen aquests elements a l'espai.

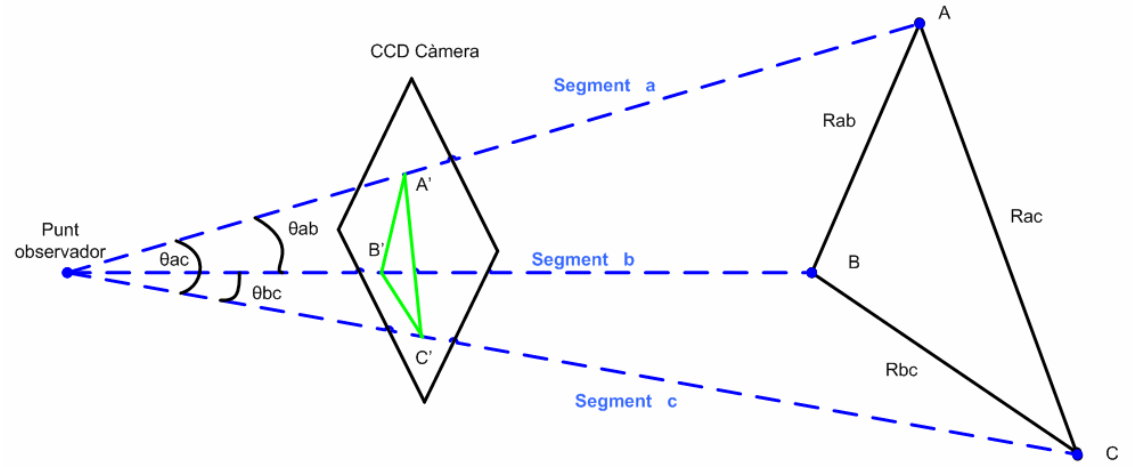


Figura 9.9 Definició del problema P3P, on tres punts d'ubicació coneguda a l'espai A, B i C, localitzats a la imatge en A', B' i C' permetran ubicar la càmera a l'espai. En la figura es defineixen les distàncies entre els punts de l'espai: Rab, Rac i Rbc, les distàncies dels tres punts a la càmera com a, b i c, i els angles entre les rectes del punt d'observació als punts com θ_{ab}, θ_{bc} i θ_{ac}.

Per construcció del tetràedre format per A, B, C i la càmera es plantegen les equacions per localitzar la càmera a l'espai (veure equació 9.3) que pel mètode de substitució acaben en un polinomi de grau quatre. Les quatre solucions del polinomi es

poden classificar fàcilment [Gao 03] per seleccionar la correcta i resoldre la ubicació de la càmera.

$$Rab = \sqrt{a^2 + b^2 - 2 \cdot a \cdot b \cdot \cos\theta_{ab}}\ ;$$
$$Rbc = \sqrt{b^2 + c^2 - 2 \cdot b \cdot c \cdot \cos\theta_{bc}}\ ;$$
$$Rac = \sqrt{a^2 + c^2 - 2 \cdot a \cdot c \cdot \cos\theta_{ac}} \qquad \text{(Eq. 9.3)}$$

Per trobar els punts A, B i C a la imatge s'han situat damunt uns superfície de l'escena unes marques, que són seguides per un algoritme de *tracking*. També s'ha situat una quarta marca, anomenada punt D, per si s'escau, poder reduir els errors en la resolució del problema. La taula 9.8 mostra els temps d'execució de la solució del problema P3P que caldrà sumar al temps d'obtenció de les vistes en el sistema interactiu de realitat augmentada.

Subtasca de solució del problema P3P	**Temps d'execució en PC**
Captura de la imatge	5ms
Identificació dels punts A, B i C	4ms
Resolució del sistema d'equacions	12ms
Determinació de la ubicació de la càmera	3ms
Temps total de la solució	**24ms**

Taula 9.8 Temps d'execució dels procés de localització de la càmera a l'espai a partir de tres punts.

Un cop localitzada la càmera a l'espai, es defineix a partir d'A, B i C un nou sistema de coordenades XYZ (veure figura 9.10) que es farà coincidir amb el sistema de coordenades del món virtual, i ja es podrà definir un punt o àrea d'inserció dels objectes.

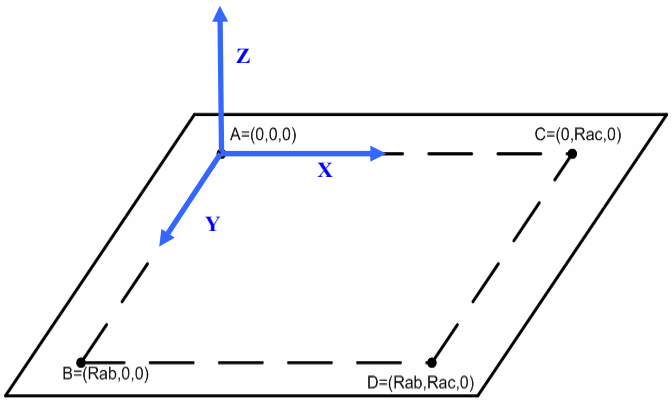
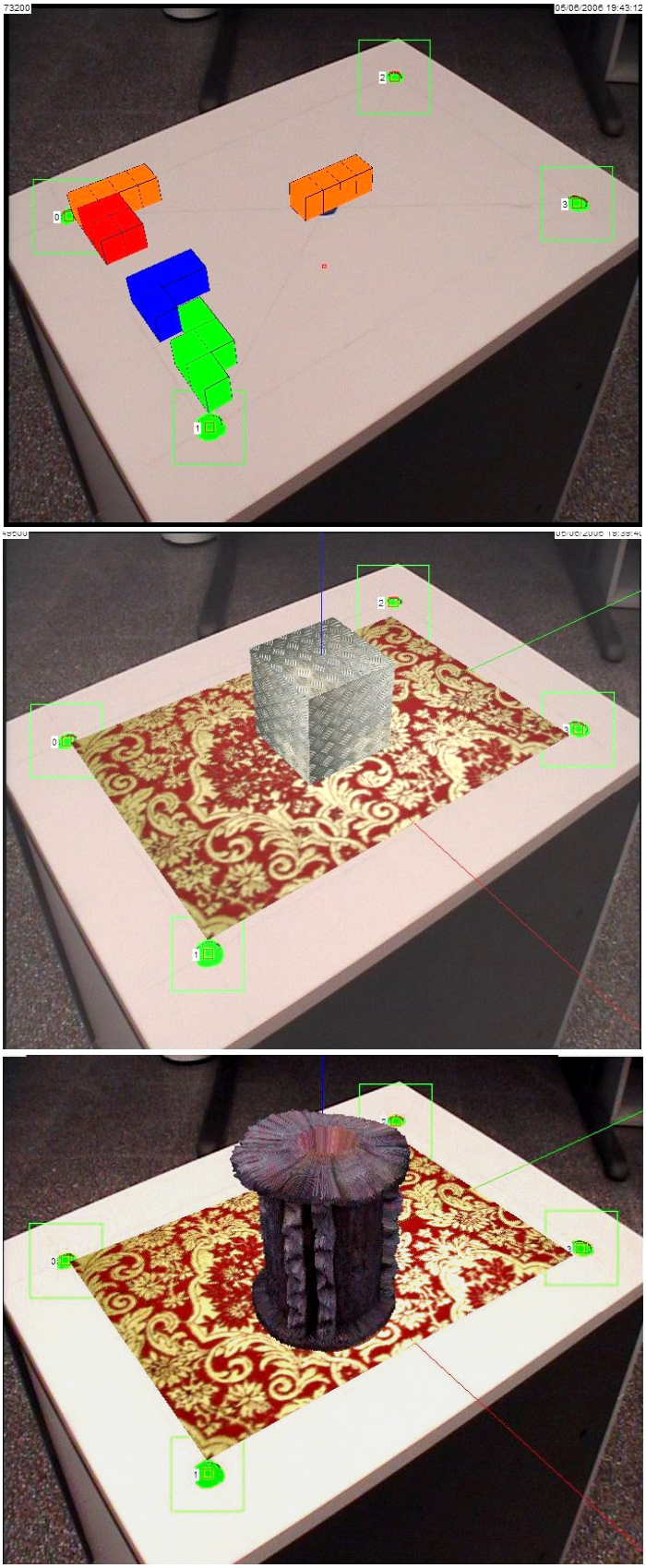


Figura 9.10 Definició del nou sistema de coordenades damunt les marques per la correspondència entre el món real i el virtual.

Sols queda mostrar els resultats dels experiments en realitat augmentada. La figura 9.11 mostra la implementació del sistema damunt una taula amb marques. Primer s'ha implementat un joc de *tetris,* després una plataforma per mostrar objectes virtuals i finalment s'han substituït els objectes virtuals per objectes descrits amb vistes segons les tècniques exposades en aquesta tesi.



(a)

(b)

(c)

Figura 9.11 Exemples del sistema de realitat augmentada, on es fa el seguiment de quatre marques (en verd) per resoldre la ubicació de la càmera. El primer mostra un joc de *tetris* damunt una taula on els objectes no tenen textura, el segon un objecte virtual on s'ha aplicat una textura metal·litzada sintètica i el tercer exemple, mostra vistes d'un objecte real on la seva informació de textura ha estat capturada amb el sistema d'adquisició robotitzat.

9.2.2 Sistema de realitat augmentada amb vistes d'objectes reals.

Com s'ha vist en els exemples de l'apartat 9.2.1, s'ha implementat un sistema de realitat augmentada on, en comptes de línies, punts o xifres, es mostren vistes d'objectes reals en correspondència amb el món real. Sistemes com el mostrat hauran d'emmotllar-se a una estructura genèrica per aquesta mena d'aplicacions, que es mostra en la figura 9.12. El sensor que permet ubicar la càmera real a l'espai pot ser, com en el cas presentat, la mateixa càmera. L'usuari seleccionarà el punt d'inserció de l'objecte virtual, el que permetrà, juntament amb la ubicació de la càmera, obtenir la vista requerida de l'objecte. La posada en correspondència permetrà alinear correctament el món real i el virtual, de manera que podran ser sumats per poder-se mostrar a l'usuari.

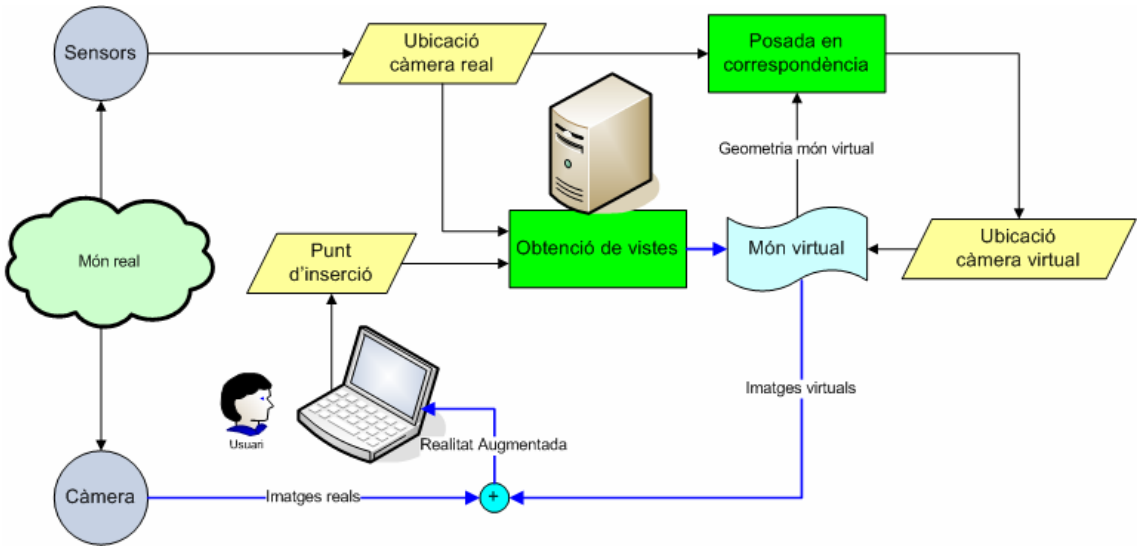


Figura 9.12 Diagrama de blocs per un sistema genèric de realitat augmentada amb vistes d'objectes reals.

Com a sistema d'obtenció de vistes podrà usar-se qualsevol dels mostrats al capítol setè, tots ells necessiten la posició de l'observador i el punt d'inserció per calcular els angles de selecció de la vista. En funció de la tecnologia triada es podrà arribar a fer la inserció d'un objecte a *video-rate*.

L'esquema mostrat és per la inserció de vistes d'un objecte real damunt de les imatges proporcionades per una càmera. La rèplica del sistema d'adquisició o càmera, permetrà treballar amb un sistema de realitat augmentada estereoscòpic. La selecció de múltiples punts d'inserció per diferents objectes, permetrà crear tota una escena virtual en conjunció amb la real.

9.3 Aplicacions en telepresència

La segona aplicació donada als mètodes d'obtenció de vistes és la de mostrar remotament objectes a través d'internet. La idea és senzilla, ja que consisteix només en anar mostrant les vistes damunt la pantalla d'un terminal remot. La tria del terme telepresència no ha estat fàcil, ja que comunament s'entén com a desplaçament de l'observador cap a una altra ubicació, quan en aquest cas el que es desplaça són les

vistes d'un objecte. L'ús del terme televisió, massa associat al dispositiu comercial s'ha considerat menys adequat, mentre que el de teleoperació, tot i que l'objecte pot ser mogut per l'usuari, tampoc s'ha triat perquè sol associar-se a operació remota de robots.

S'està treballant en un sistema de presentació remota de vistes d'objectes juntament amb el departament d'arqueologia de la Universitat de Barcelona (projecte AISCA[1]), per mostrar les vistes dels objectes en un museu a internet. En funció del mètode triat d'obtenció de vistes, caldrà plantejar d'una manera o altra l'arquitectura del sistema. A continuació es mostra, per cada un dels mètodes presentats, una proposta d'implementació del sistema i una avaluació del seu funcionament segons experiments realitzats.

9.3.1 Sistema de visualització remota d'objectes pel mètode d'accés a vistes.

El primer sistema avaluat és el de disposar d'un servidor de vistes al que es connecten els usuaris per demanar interactivament vistes dels objectes. En aquest cas l'usuari farà peticions de vistes d'un objecte que hauran de ser respostes per la base de dades present al servidor (figura 9.13).

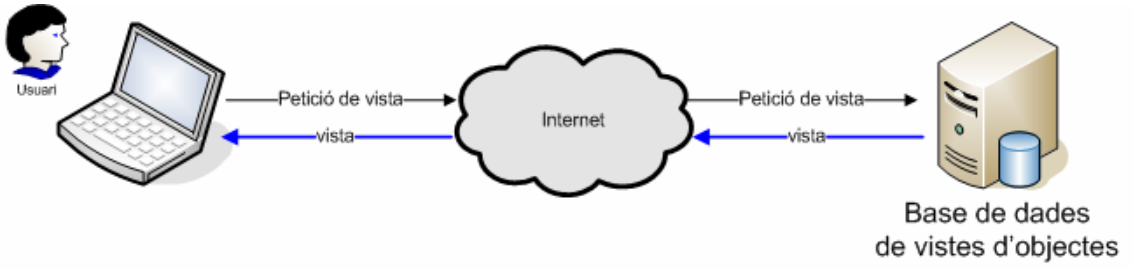


Figura 9.13 Diagrama de blocs d'un sistema de visualització remota a través d'internet per accés a una base de dades de vistes.

Els principals problemes d'aquesta arquitectura seran tres: per una banda, el servidor de base de dades de vistes haurà de proporcionar qualsevol vista al client o clients, per altra banda, per complir les condicions d'interactivitat, la xarxa haurà de garantir el lliurament d'una vista en un temps determinat i finalment, serà necessari que el client disposi d'un descompressor d'imatges suficientment ràpid.

- El primer problema, es pot resoldre fent que el servidor tingui a memòria totes les vistes de l'objecte: en l'experimentació feta un objecte té 1200 per 300 vistes que en qualitat JPEG ocupen cada una d'elles uns 10KB a memòria, és a dir amb 4GB de memòria es poden tenir totes les vistes carregades.
- Pel segon problema fa falta gaudir de l'ample de banda necessari, i el que és més important, l'anomenada "qualitat de servei" que ha d'acotar el temps de lliurament dels paquets de dades. Actualment, qualsevol connexió tipus ADSL pot garantir un ample de banda suficient (entorn 1-2Mbps) però sols les xarxes ATM garanteixen la qualitat de servei. Les aplicacions de vídeo sota demanda damunt protocols com TCP/IP basen el seu funcionament en el "*buffering*" que realitza l'equip de visualització. En el cas presentat, un "*buffering*" aniria contra la interactivitat del sistema i sols es podria fer en base a enviar paquets grans de

[1] AISCA és l'acrònim de Archaeological Information System of Central Asia

vistes, de manera que per la vista de longitud (i, j) s'enviessin les 8 o 24 vistes veïnes, preveient les possibles peticions de vistes dels usuaris.

- Per aconseguir una descompressió ràpida de les vistes, caldrà emprar el suport que el maquinari dels processadors moderns ofereix a les tasques de tractament d'imatge. De totes maneres, la descompressió JPEG està pensada per ser usada sota demanda de l'usuari i ofereix un rendiment, d'unes 8 o 9 vistes per segon.

En els experiments fets, el procés d'accés a vistes a través d'internet ha donat un rendiment màxim de tres vistes per segon, limitat pel temps de descompressió. Caldrà investigar més en com usar el suport del maquinari (pensat per descompressió de seqüències de vídeo) per la descompressió de vistes similars, però que arriben de manera discreta i no seqüencial.

9.3.2 Sistema de visualització remota d'objectes per representació de models tridimensionals.

En aquesta opció, l'aplicació client haurà de demanar en primer lloc, tota la informació relativa a l'estructura tridimensional de l'objecte i un cop rebuda, podrà treballar localment mitjançant la projecció del model amb el seu coprocessador gràfic o GPU.

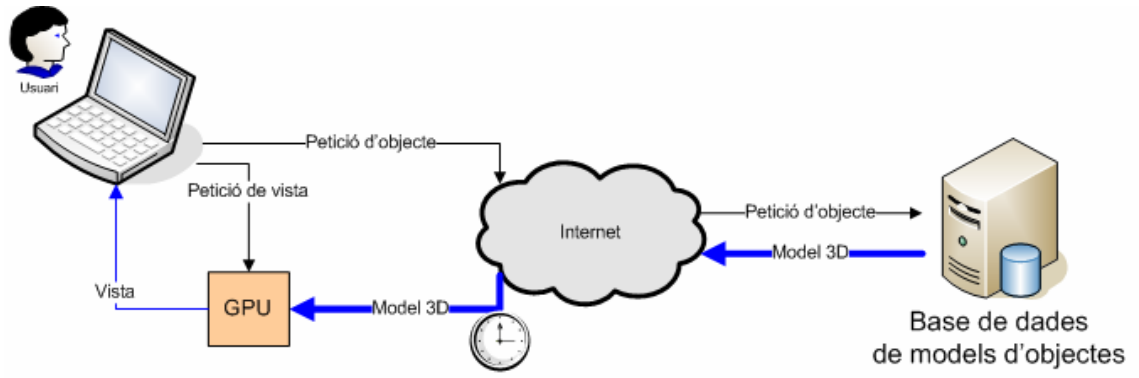


Figura 9.14 Diagrama de blocs d'un sistema de representació de models tridimensionals amb textura, obtinguts d'una base de dades d'objectes.

Les principals restriccions temporals d'aquest mètode seran doncs, el temps de càrrega de les dades (per un fitxer com els obtinguts i amb una línia ADSL convencional al voltant d'un minut) i un cop l'objecte es troba en el seu equip el temps de projecció de l'escena tridimensional del coprocessador gràfic. En els experiments realitzats per objectes d'entorn d'un milió de triangles s'han obtingut cinc vistes per segon, tot i que s'espera que amb millores en la programació i en el maquinari es pugui arribar fàcilment a vint o trenta imatges projectades per segon amb les que interactuar.

9.3.3 Sistema de visualització remota d'objectes pel mètode de síntesi de vistes.

Aquesta tercera arquitectura plantejada, amb la utilització del mètode de síntesi de vistes ofereix moltíssimes similituds amb la opció anterior. Observant les figures 9.14 i 9.15 es pot veure que, en ambdós casos cal realitzar una petició de les dades de

l'objecte, en aquest cas un conjunt de vistes i mapes de correspondències i posteriorment, ja en l'equip local, realitzar la visualització de la informació. En aquest cas, la càrrega del procés recau al processador principal de l'equip, que serà l'encarregat de realitzar les reprojeccions, interpolació i presentació de les noves vistes de l'objecte sota demanda de l'usuari.

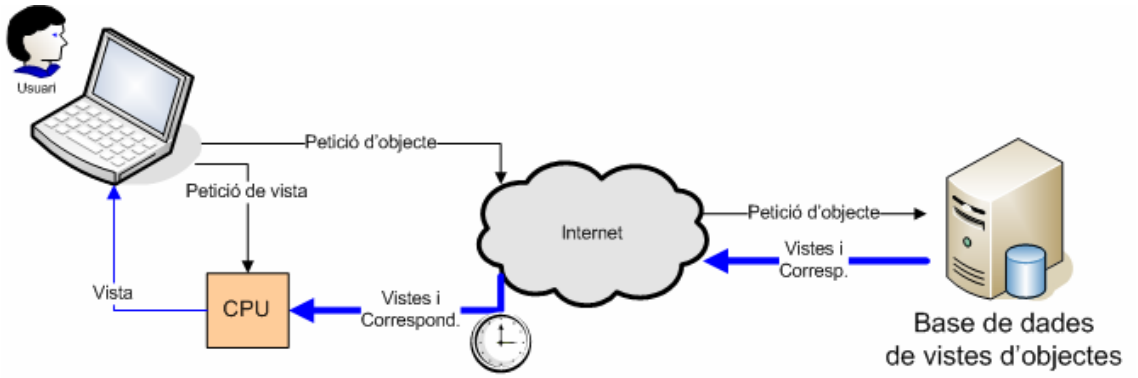


Figura 9.15 Diagrama de blocs del sistema d'obtenció de vistes reals d'un objecte a partir de vistes seleccionades, mapes de correspondència i un algoritme d'interpolació executat localment a la CPU de l'equip client.

El rendiment obtingut per aquest sistema és similar a l'anterior i, un cop descarregada la informació de l'objecte (al voltant d'un minut també), depèn de la capacitat de càlcul del processador. En els experiments fets un màxim de cinc vistes per segon amb expectatives de millorar properament.

10. Conclusions

En aquesta tesi s'han realitzat una sèrie d'experiments per tal de cercar, identificar, caracteritzar i comparar diversos mètodes d'obtenció de vistes d'objectes reals per aplicacions de realitat augmentada, telepresència o altres que puguin idear-se en el futur. Malgrat que els experiments sols han estat això, experiments, i els seus resultats numèrics poden estar esbiaixats per les implementacions realitzades, si que han servit per emmarcar els mètodes trobats en el context tecnològic actual, trobar-ne els punts forts i febles i preveure'n l'evolució i aplicabilitat.

Durant el desenvolupament dels mètodes trobats, de naturalesa diversa, han anat sorgint problemes i curiositats que han fet aprofundir aquest treball en l'àmbit de la geometria de la síntesi de vistes, la reconstrucció de l'estructura tridimensional dels objectes, l'acceleració de certs algoritmes amb l'ajut del maquinari existent o la portabilitat de les dades a través de la xarxa internet. Tot i iniciar la tesi amb certs prejudicis a favor de les tècniques d'interpolació de vistes, la concreció del mètode d'accés a vistes preenregistrades i el d'aplicació de textures al model tridimensional, els ha fet avaluar en igualtat amb l'inicial, trobant que cada un pot tenir el seu escenari d'aplicació. A continuació es mostraran els objectius assolits per la tesi, les aportacions que s'han realitzat, unes propostes de treball futur i unes reflexions finals sobre els mètodes presentats amb una previsió de l'evolució de la tecnologia involucrada en la seva implementació.

10.1 Objectius assolits

Dels objectius plantejats a l'inici d'aquesta tesi, es considera que s'han complert els principals, tot i quedar algunes tasques per realitzar que, juntament, amb d'altres que s'han plantejat, es proposaran com a treball futur.

- S'han identificat, definit, provat i avaluat diversos mètodes d'obtenció de vistes d'objectes reals amb possibilitat d'aplicació en entorns de realitat augmentada o presentació remota de vistes d'objectes.
- Pel mètode de síntesi de vistes a partir d'un subconjunt inicial i tècniques d'interpolació, s'ha definit un procediment de selecció de vistes, gravació de la informació mínima necessària per a la reconstrucció, reprojecció de vistes per a tenir les condicions de geometria epipolar i interpolació de la nova vista de l'objecte.

- S'han experimentat diversos mètodes d'obtenció de la informació tridimensional necessària per a la síntesi de vistes, proposant millores en les implementacions d'alguns d'ells.
- S'ha obtingut, amb l'ajuda d'una plataforma robotitzada creada per als experiments, un conjunt de dades per a l'experimentació que ha permès l'avaluació d'errors, comparació de mètodes i definició de nous mètodes.
- S'ha experimentat la introducció de les vistes obtingudes en un entorn de realitat augmentada interactiu, i durant el desenvolupament de la tesi, s'ha concretat un experiment de visualització remota d'objectes arqueològics.
- Per tots els mètodes avaluats s'han cercat les seves restriccions temporals, les seves necessitats de recursos dels ordinadors i les principals fonts d'errors en la seva realització.
- S'han emprat criteris diversos de comparació de la qualitat i la fidelitat de les imatges obtingudes, des del criteri subjectiu (lògic tenint en compte que el destinatari de la informació és un humà) a d'altres de numèrics, que permeten l'ús de la mesura d'error com a element de control en la generació de les vistes.

10.2 Treball futur

Han quedat algunes tasques pendents d'execució, no per poc importants, sinó perquè algunes d'elles necessitarien potser un altre treball de tesi complert per ser realitzades, d'altres queden fora de l'àmbit del coneixement de l'autor i un altre grup serien només orientades al lluïment en el sentit més tecnològic, però sense cap aportació especial. En aquelles que tenen més interès científic s'espera continuar-hi treballant, ja que obren camins d'investigació interessants. Com a treball futur sorgit del fet en aquesta tesi es planteja:

- Estudiar els fenòmens derivats de la inserció de les vistes dels objectes reals en l'entorn de realitat augmentada: el cosit òptim de dues imatges preses per càmeres diferents, amb paràmetres diferents, il·luminacions diferents i potser escales diferents.
- Augmentar la resolució del volum de vòxels necessari en la reconstrucció per *space carving* i la precisió en aquest procediment, fins permetre que el mètode d'aplicació de textures al model arribi a tenir una qualitat similar als altres. Paral·lelament, seguir treballant en l'acceleració del procés de reconstrucció i minimització de les dades necessàries per a la representació de l'objecte.
- Demostrar matemàticament els criteris proposats de selecció de vistes, que han donat bon resultat experimental, o proposar uns altres criteris millors o més complerts, per trobar el nombre mínim de vistes representatiu d'un objecte.
- Realitzar millores en la implementació de la part interactiva del mètode de síntesi de vistes, en la programació del coprocessador gràfic pel mètode de projecció de models i en la organització de la informació gravada en disc i memòria pel mètode de selecció de vistes. D'aquesta manera es podria arribar a rendiments propers a l'anomenat *video-rate* que farien percebre els moviments de l'objecte de forma contínua a l'usuari

10.3 Aportacions

Durant la realització de la tesi han anat sorgint idees, petites millores de mètodes existents o concrecions d'algoritmes nous que representen petites aportacions en el seu camp de coneixement. A continuació es mostrarà una llista d'aquestes aportacions; algunes han estat referendades en forma de publicacions, altres estan pendents de publicar en la data de publicació d'aquest volum. El camí seguit en l'elaboració de la tesi ha portat a publicar en diversos àmbits, sempre en l'àrea plantejada i, el fet de ser un treball realitzat per enginyers, en un departament d'enginyeria i en una universitat politècnica ha fet que l'enfocament sigui, en general, més aplicat que científic. Les aportacions realitzades són:

- Disseny d'una metodologia per la representació d'objectes a partir de conjunts de vistes i mètodes de síntesi. En aquesta metodologia s'ha presentat un protocol per l'adquisició i ordenació de les dades, idees per la selecció del conjunt mínim de vistes, un criteri per gravar la mínima informació necessària, ajuts a l'obtenció de la informació tridimensional de l'escena necessària, i un algoritme ràpid i general de síntesi de vistes.
- Identificació i caracterització dels tres mètodes per a l'obtenció de vistes d'objectes reals. Comparació dels tres mètodes, amb la selecció de mesures de la fidelitat de les vistes a l'original, informació mínima necessària pel seu funcionament i temps de còmput.
- Supressió de restriccions geomètriques del mètode de síntesi per rectificació de tres vistes, permetent generalitzar la ubicació de la càmera virtual [Martín 03], publicat a LNCS[1]. En el mateix treball es presentava una millora en el càlcul de la distància del pla de reprojecció per maximitzar l'àrea de la vista interpolada. El capítol quart d'aquesta tesi reflexa aquest treball.
- Especificació de l'algoritme de síntesi de vistes pel mètode de rectificació de tres vistes, de forma que es pugui implementar amb processadors vectorials, eines com DSP o conjunts d'instruccions específiques (extensions multimèdia) dels processadors CISC, per assolir les necessitats de les aplicacions interactives.
- Presentació d'un mètode de refinament de models tridimensionals obtinguts per *space carving* mitjançant estereovisió. El mètode combina dues tècniques conegudes de visió per ordinador obtenint un millor resultat en la reconstrucció tridimensional [Martín-Aranda 03]. S'ha presentat en el capítol sisè d'aquesta tesi.
- Acceleració del mètode de reconstrucció tridimensional per projecció de vòxels amb la utilització de mapes de distància, estructures en arbre i el coprocessador gràfic present en els computadors personals, tal com es mostra en el capítol sisè de la tesi.
- Discussió de l'ús de les vistes d'objectes reals en realitat augmentada en entorns industrials [Martín 01] i de jocs d'ordinador [Arboleda 02].

[1] LNCS és l'acrònim de Lecture Notes on Computer Science.

- Translació dels mètodes d'accés a vistes d'objectes per aplicació a bases de dades d'objectes arqueològics, amb la intenció de mostrar remotament les vistes dels objectes reals juntament amb altra informació relativa al seu lloc de trobada, marc històric, ubicació actual, etc. (projecte AISCA, mencionat a l'apartat de resultats)

10.4 Reflexions finals

Per acabar aquesta tesi, es creu necessari fer una reflexió sobre les perspectives dels tres mètodes d'obtenció de vistes presentats. Cada un d'ells té una naturalesa prou diferent que fan que la seva comparació hagi estat dificultosa durant tota la tesi; un d'ells es basa en l'ocupació de disc, l'altre en l'ús del coprocessador gràfic i el tercer en l'ús del processador principal per calcular noves vistes.

El mètode d'accés a vistes indexades és a priori, consumidor de memòria. Una descripció completa d'un objecte pot ocupar, sense compressió, uns centenars de gigabytes i amb compressió uns centenars de megabytes. A mida que es va traient volum de dades de memòria, augmentarà el cost algorísmic de la descompressió. Donat que existeix un interès comercial al respecte, en els darrers anys ha augmentat la qualitat dels mètodes de compressió i descompressió de vídeo i el suport dels ordinadors a aquests mètodes, amb instruccions dedicades i coprocessadors integrats en el maquinari. De totes maneres, aquest suport està pensat per la descompressió de seqüències d'imatges (pel·lícules) i no toleren els salts de seqüència plantejats en el mètode d'accés a vistes. Per aprofitar aquests recursos caldrà fer un esforç en la codificació de les vistes per fer-la compatible amb els formats de vídeo. Es pot, per exemple, agrupar paquets d'imatges de vistes veïnes (en sentit horitzontal i vertical) que es podran descomprimir fàcilment i guardar en la memòria principal fent un efecte de memòria cau. Així es reduirà la dependència de l'espai a disc de les vistes tot permetent la interactivitat. Per altra banda, la mida de la memòria principal dels equips, la capacitat d'emmagatzemament dels discs i la seva velocitat de transferència també augmenta any a any. Malgrat tot, no augmenta prou ràpid com per fer pensar que es podran guardar descripcions de diversos objectes sense recórrer a les tècniques de compressió.

Així doncs, pel mètode d'accés a vistes preenregistrades, tot fa pensar que el compromís entre quantitat de procés i volum de dades es decantarà cap al primer, i la CPU s'encarregarà d'anar col·locant a la memòria principal les vistes de l'objecte. L'augment de la capacitat de memòria principal permetrà augmentar el nombre d'objectes descomprimits alhora (amb la creació d'escenes complexes) i l'augment de capacitats del disc el volum de les bases de dades. El millor d'aquest mètode és la seva senzillesa, sols cal un robot posicionador per prendre les vistes de l'objecte i espai a disc per emmagatzemar-les. En el moment actual, a partir d'un cert factor de compressió, és el mètode que ofereix millors resultats de cara a l'usuari. També és l'únic mètode que permet la visualització remota sense que l'usuari hagi de portar les dades a l'equip local. Amb la tecnologia actual es pot implementar ja un servidor de vistes d'objectes a través d'internet, sols cal que el canal emprat (servidors, encaminadors i proveïdors) garanteixin la qualitat del servei: ample de banda i temps de lliurament dels paquets de dades. No és habitual però si és possible.

El mètode de projecció de models tridimensionals amb textura, és el mètode que, en opinió de l'autor, s'imposarà a mig termini. Els coprocessadors gràfics (comunament anomenats VGA de l'antic concepte *video grafic adapter*) augmenten ràpidament la seva capacitat de projecció de triangles amb textura, impulsats pel mercat dels videojocs. No sols duen cada cop processadors més ràpids, amb arquitectures dedicades, sinó que a diferència de en altres tipus de dispositius, s'imposa la paral·lelització: actualment incorporen fins a trenta-dos línies d'execució. A més d'aquesta estructura paral·lela interna, s'han creat bussos com el *PCI-Express* que permeten la connexió de diversos dispositius per sumar les seves capacitats. Això fa que (suportant el cost econòmic, el consum elèctric i el problema de dissipació de temperatura), els usuaris no dubten en instal·lar en equips domèstics capacitats de 500 milions de triangles projectats per segon. Els creadors d'infografia estimen que, a partir d'uns 10.000 milions de triangles projectats per segon, l'ull humà no podrà distingir objectes reals d'objectes virtuals. En el cas d'objectes amb textures reals, com és el cas, aquest límit pot ser inferior. Com s'ha vist, el model tridimensional d'un objecte d'uns 50 milions de triangles ocuparà uns 80-90 MB al disc, que és un volum reduït. Això fa pensar que, en pocs anys, o actualment amb tècniques de programació de gràfics millors que les emprades, es podria arribar a projectar vistes de l'objecte real reconstruït tridimensionalment amb velocitat suficient.

El problema queda doncs circumscrit a l'habilitat de generar un model suficientment precís de l'objecte i és evident que un sistema robotitzat comercial oferirà aquesta prestació (s'està parlant de que un objecte cúbic de 20 centímetres de costat caldrà reconstruir-lo amb talls d'una dècima de mil·límetre). De cara a aplicacions de visualització remota sols hi haurà el problema de l'arrancada, ja que la descàrrega d'un fitxer de 90MB trigarà entre 2 i 20 minuts en funció de l'ample de banda disponible per l'usuari (segons els estàndards disponibles la primavera de 2006).

Finalment, el mètode de síntesi de vistes és el que, de moment, ofereix millor qualitat en la imatge obtinguda en condicions d'igualtat del volum de dades disponible. També te l'avantatge de que no és necessari ni fer la reconstrucció tridimensional completa de l'objecte, ni adquirir-ne totes les vistes (no tots els objectes són susceptibles de ser col·locats en un robot posicionador). A partir d'algunes vistes i informació tridimensional (que pot no explicitar-se) obtinguda per esculpit de vòxels i/o tècniques estereoscòpiques es poden interpolar totes les altres. El mètode de síntesi de vistes garanteix que, si es té un mapa de disparitat dens, no hi ha oclusions ni forats, pot generar qualsevol vista físicament correcta. Desgraciadament, aquestes condicions no es donaran sempre i llavors la qualitat de la vista obtinguda se'n resenteix. L'altre punt feble del mètode és la dependència única del processador per realitzar els càlculs necessaris. S'estima que, programant adequadament les instruccions vectorials del processador, agrupades en les MMX (acrònim de *MultiMedia eXtensions*) s'assolirà en breu un rendiment de vint o trenta imatges generades per segon. En qualsevol cas, a mida que es vulguin introduir més objectes a l'escena, apareixerà de nou el coll d'ampolla del temps de procés. Per aplicacions de visualització remota d'objectes, hi ha l'inconvenient del temps de descàrrega de les dades de l'objecte i la saturació del processador de l'equip local.

Per tot l'exposat, no es pot concloure que un dels mètodes sigui millor o pitjor que els altres. Com s'ha dit, el mètode de selecció i síntesi de vistes és el que, a data d'avui, pot oferir millor qualitat d'imatge amb menor volum de dades emmagatzemades i té l'avantatge de no requerir necessàriament l'adquisició de totes les vistes o reconstruir amb molta precisió el model tridimensional i per això s'hi ha treballat especialment en aquesta tesi. Per aplicacions interactives de realitat augmentada o visualització remota amb diversos objectes, el mercat imposarà la utilització del coprocessador gràfic i els models amb textura. Finalment, per la visualització remota a través de la xarxa de pocs objectes, amb interès de preservar la base de dades de vistes i un grau d'interactivitat limitat (com és el cas del treball plantejat amb peces arqueològiques), el mètode d'accés a vistes pot oferir les millors prestacions.

Bibliografia

Part 1. Referències clàssiques

S'inclouen aquí les citacions de llibres, articles en revista o congrés, tesis doctorals i reports de recerca d'empreses o universitats.

[Adelson-Bergen 91] E.H. Adelson, J.R. Bergen, *The Plenoptic Function and the Elements of Early Vision*, Computational Models of Visual Processing, Capítol 1. MIT-Press, Cambridge, Massachusetts. 1991.

[Ahumada 93] A. J. Ahumada Jr., *Computational image quality metrics: a review*. Research Report. NASA Ames Research Center. 1993.

[Arboleda 02] Juan P. Arboleda, A.B. Martínez, E.X. Martín, *Mixed Reality in traffic Scenes. A view synthesis approach*. International Woekshop on Entertainment Computing. Tokyo, 2002.

[Azuma 97] R. Azuma, *A Survey of Augmented Reality*. Presence: Teleoperators and Virtual Environments vol.6, pàgines 355-385. 1997.

[Azuma 99] Ronald T. Azuma, HRL Laboratories, USA, *The Challenge of Making Reality Work Outdoors*, del llibre *Mixed Reality, merging Real and Virtual Worlds,* editat per Y. Ohta, H. Tamura, Ed. Ohmsha, Ltd. Tokyo. 1999.

[Baker 99] S. Baker, R. Szeliski, P. Anandan. *A layered approach to stereo reconstruction.* Proceedings Computer Vision and Pattern Recognition Conference, pàgines 434–441. 1998.

[Beier-Neely 92] Thaddeus Beier, Shawn Neely, *Feature-Based Image Metamorphosi*", Proceedings SIGGRAPH'92, pàgines 35-42. 1992.

[Canon 95] Akihiro Katayama, Koichiro Tanaka, Takahiro Oshino, Hideyuki Tamura. Media Technology Laboratory, Canon Inc. Kawasaki, Japan. *A viewpoint dependent stereoscopic display using interpolation of multi-viewpoint images,* Proceedings SPIE Stereoscopic Displays and Virtual Reality Systems II, vol. 2049, pàgines 11-20. 1995.

[Chen-Williams 93] S.E. Chen, L. Williams, "View Interpolation for Image Synthesis", Computer Graphics, Proceedings SIGGRAPH 93, pp. 279-288, Juliol 1993.

[Chien 86] C. H. Chien and J. K. Aggarwal. Volume surface octrees for the representation of 3D objects. *Computer Vision, Graphics and Image Processing*, número 36, pàgines 100–113, 1986.

[Collignon 95] A. Collignon, F. Maes et al. *Automatic multimodality image registration using information theory.* Proceedings of the International Conference on Information Processing Medical Imaging. IPMI, pàgines 263-364. 1995.

[Connolly 85] C. Connolly, *The determination of next best views.* Proceedings of IEEE International Conference on Robotics and Automation, pàgines 432-435. 1985.

[Conan-Doyle 1891] Sir Arthur Conan-Doyle, *A scandal in Bohemia*. 1891.

[Delso 03] G. Delso. *Registro Elástico de Imágenes Médicas Multimodales*. Tesi doctoral Universitat Politècnica de Catalunya, dept. CREB-ESAII. 2003.

[Faugeras 93] O. Faugeras, *Three dimensional computing vision.* MIT Press, Cambridge, Massachusetts, 1993.

[Fau-Laveau-Robert 95] Olivier Faugeras, Stéphane Laveau, Luc Robert, *3-D Reconstruction on Urban Scenes from Sequences of Images*, Report de Recerca INRIA núm. 2572. Juny 1995.

[Faugueras-Keriven 96] Olivier Faugueras (INRIA-MIT), Renaud Keriven (ENPC), *Variational principles, Surface Evolution, PDE's, level set methods and the Stereo Problem*, Report de Recerca INRIA núm. 3021, 26 Octubre 1996.

[Faugeras-Robert 93] Olivier Faugueras, Luc Robert, *What can two images tell us about a third one?*, Report de Recerca INRIA núm. 2018, Juliol 1993. Publicat també com a:

[Faugeras 96] O. Faugeras, L. Robert, *What can two images tell us about a third one?* International Journal on Computer Vision, número 18, pàgines 5-19. 1996.

[Fischler 81] Martin A. Fischler, Robert C. Bolles, *Random Sample Consensus: A paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*. Communications of the ACM, Juny 1981, número 6, volum 24.

[Forsyth 02] David A. Forsyth, Jean Ponce, *Computer Vision, a Modern Approach*, Editorial Prentice Hall, 2002.

[Fuiji-Harashima 94] T. Fuiji, H. Harashima, The University of Tokyo*, Coding of an autostereoscopic 3-D image sequence*, Proceedings SPIE Visual Communication and Image Processing, (VCIP'94), vol. 2308, pàgines 930-941. 1994.

[Fuchs_99] Henry Fuchs, Jeremy Ackermann. *Displays for Augmented Reality: Historical Remarks and Future Prospects,* Mixed Reality - Merging Real and Virtual Worlds, Ohmsha (Tokyio)-Springer Verlag (Berlin), pàgines 31-40. 1999.

[Fuchs-Ackerman 99] Henry Fuchs, Jeremy Ackerman, The University of North Carolina at Chapel Hill, USA, *Displays for Augmented Reality: Historical Remarks and Future Prospects*, del llibre *Mixed Reality, merging Real and Virtual Worlds,* editat per Y. Ohta, H. Tamura, Ed. Ohmsha, Ltd. Tokyo. 1999.

[Gao 03] Xiao-Shan Gao, Xiao-Rong Hou et al. *Complete Solution Classification for the Persepective-Three-Point Problem*, IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 25, numero 8, Agost 2003.

[Hill 94] D.L.G Hill, D.J. Hawkes. *Voxel similarity measures for automated image registration.* Proceedings Visualization for Medical Computing. Pàgines 205-216. 1994.

[Hirose 99] Michitaka Hirose, Tomohiro Tanikawa, The University of Tokyo, Takaaki Endo, Mixed Reality Systems Laboratory, Japan, *Building a Virtual World from the Real World*, del llibre *Mixed Reality, merging Real and Virtual Worlds,* editat per Y. Ohta, H. Tamura, Ed. Ohmsha, Ltd. Tokyo 1999.

[Huang 98] Ho-Chao Huang, Ching-Che Kao, Yi-Ping Hung, Shung-Hua Nain, Institute of Information Science, National Taiwan University, Taipei, Taiwan. *Generation of Multiple perspective Videos from Two Views*, Proceedings Int. Computer Simposium, 1998.

[Jain 95] Ramesh Jain, R. Kasturi, B. Schunck, *Machine Vision,* Editorial McGraw-Hill. 1995.

[Kanade-Fuchs 94] Takeo Kanade, Carnegie Mellon University, Henry Fuchs, Gary Bishop, University of North Carolina, *Virtual Space Teleconferencing using a Sea of Cameras*, UNC-Reports, 1994.

[Kanade_96] T. Kanade. *A Video-Rate Stereo Machine and Its New Applications*, Computer Vision and Pattern Recognition Conference, 1996, San Francisco, CA

[Kimura 99] M. Kimura, H. Saito, and T. Kanade. *3D voxel construction based on epipolar geometry*. Proceedings . Int. Conf. Image Processing, pàgines 135–139. 1999.

[Kutulakos 00] K. Kutulakos and S. Seitz, *A theory of shape by space carving.* International Journal on Computer Vision, 38, pàgines 198-218, any 2000.

[Laurentini 94] A. Laurentini, *The Visual Hull Concept for Silhouette-Based Image Understanding.* IEEE Transactions on Pattern Recognition and Machine Intelligence, 16, pàgines 150-162, Febrer 1994.

[Laveau-Faugueras 94] Stéphane Laveau, Olivier Faugueras, *3-D Scene Representation as a Collection of Images and Fundamental Matrice*s, Report de Recerca INRIA núm. 2205, Febrer 1994.

[Laveau Faugueras 97] Stéphane Laveau, Olivier Faugueras, *Oriented projective geometry for Computer Vison*, INRIA Sophia-Antipolis, 1997.

[Lei-Hendriks 02] B.J.Lei, E.A. Hendriks, *A real-time realization of Geometrical Valid View Synthesis for Tele-conferencing with Viewpoint Adaptation*, Proceedings of First International Symposium of 3D Data Processing, Visualization and Transmission, pàgines 327-331, any 2002.

[Lhuillier 99] Maxime Lhuillier, *Towards Automatic Interpolation for Real and Distant Image Pairs*, Report de Recerca INRIA núm. 3619, Febrer 1995.

[Martín 01] E.X. Martín, A.B. Martínez, *Generation of synthetic views for teleoperation in industrial processes*. Proceedings of the IEEE International conference on Emerging Technology and Factory Automation, pàgines 537-541. 2001.

[Martín 03] E.X. Martín, A.B. Martínez, J. Aranda, *Generalising the virtual camera pose for view synthesis*. Lecture Notes on Computer Science, 2537. Ed. Springer Verlag, pàgines 701-708. 2003.

[Martin-Aranda 03] E.X. Martín, J. Aranda, A.B. Martínez, *Refining 3D recovering by carving through view synthesis and stereovision*. Lecture Notes on Computer Science, 2652. Ed. Springer Verlag, pàgines 793-801. 2003.

[McMillan-Bishop 95] Leonard McMillan, Gary Bishop, Dept. Computer Sciences, University of North Carolina at Chapel Hill, *Plenoptic Modeling: An Image-Based Rendering System* , Proceedings of SIGGRAPH 95 (Los Angeles, Agost 1995)

[McMillan 95] Leonard McMillan, Dept. Computer Sciences, University of North Carolina at Chapel Hill*, Acquiring Immersive Virtual Environments with an Uncalibrated Camera*, UNC-Technical Report 95-006.

[Medioni 04] Gerard Medioni, Sing Bing Kang, *Emerging Topics in computer vision,* Editorial prentice Hall, juliol 2004.

[Milgram 94] Paul Milgram, F. Kishino, *A Taxonomy of Mixed Reality Visual Displays*, IEIC Transactions on Information Systems, Vol E77-D, número.12, pàgines 1321-1329 .1994.

[Milgram_99] Paul Milgram, Herman Colquhoun Jr., *A Taxonomy of Real and Virtual World Display Integration*, Mixed Reality - Merging Real and Virtual Worlds, Ohmsha (Tokyo)-Springer Verlag (Berlin), pàgines 5-30, 1999.

[Müller 02] K. Müller: *MPEG-7: Content Description of Multimedia Data*, Production Reality - Journal for Broadcast, Post-Production, Animation and Content Description, pàgines 35-37, Nov./Des. 2002

[Müller 03] K. Müller, A. Smolic: *MPEG-7: Applications for TV- and Cinema Appplications:– Multimedia Content Description Interface (MPEG-7: Anwendungen für die Film- und Fernsehtechnik – das Multimedia Content Description Interface)*, FKT- Fachzeitschrift für Fernsehen, Film und Elektronische Medien, pàgines 703-705, Des. 2003.

[Müller 05] K. Müller, A. Smolic, M. Droese, P. Voigt, and T. Wiegand: *3D Reconstruction of a Dynamic Environment with a fully Calibrated Background for Traffic Scenes, IEEE Transaction on Circuits and Systems for Video Technology*, volum 15, número 4, pàgines 538-549, Març 2005.

[Naemura-Harashima 99] Takeshi Naemura, Hiroshi Harashima, The University of Tokyo, Japan *The Ray-Based Approach to Augmented Spatial Communication and Mixed Reality*, del llibre *Mixed Reality, merging Real and Virtual Worlds,* editat per Y. Ohta, H. Tamura, Ed. Ohmsha, Ltd. Tokyo. 1999.

[Narayanan-Kanade 97] P.J. Narayanan, Res. Center for IA, Bangalore, India, Takeo Kanade, Carnegie-Mellon University, *Virtualized Reality: Constructing Time-Varying Virtual Worlds From real World Events*, Proceedings of IEEE Visualization'97, pàgines 277-283. Oct. 1997.

[Narayanan-Kanade 98] P.J. Narayanan, Res. Center for IA, Bangalore, India, Takeo Kanade, Carnegie-Mellon University*, Constructing Virtual Worlds Using Dense Stereo*, Proceedings of IEEE Intl. Conf. on Computer Vision'98, pàgines 3-10. Bombay 1998.

[Ohm-Müller 99] J.-R. Ohm, K. Müller : *Incomplete 3D - Multiview representation of video objects*, IEEE Transactions on Circuits and Systems for Video Technology, special issue on SNHC, pàgines 389-400. Març 1999

[Ohta-Tamura 99] Yuichi Ohta, Hideyuki Tamura, *Mixed Reality, merging Real and Virtual Worlds,* editat per Y. Ohta, H. Tamura, Ed. Ohmsha, Ltd. Tokyo. 1999.

[Pappas 99] Thrasyvoulos N. Pappas, Robert J. Safranek. *Perceptual Criteria for Image Quality Evaluation*. Research Report. Bell Laboratories, Lucent Technologies. 1999.

[Piekarski 02] Piekarski Wayne, Thomas B, *ARQuake: The Outdoor Augmented Reality Gaming System*, ACM Communications, Volum 45, número, 1 pàgines, 36-38, juny 2002.

[Potmesil 87] M. Potmesil, *Generating octree models of 3D objects from their silhouettes in a sequence of images*. Intl. Journal on Computer Vision, Graphics and Image Processing, número 40, pàgines 1–20. 1987.

[Saito 99] H. Saito and T. Kanade. *Shape reconstruction in projective grid space from large number of images.* In Proc. Computer Vision and PatternRecognition Conf., volume 2, pages 49–54, 1999.

[Scharstein 99] Daniel Scharstein, *View Synthesis Using Stereo Vision*. Lecture Notes on Computer Science, Editorial Springer-Verlag, número 1583. 1999.

[Scharstein 02] D. Scharstein, R. Szeliski. *A taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms.* International Journal on Computer Vision. Pàgines 7-42, Juny 2002. Microsoft Research Technical Report MSR-TR-2001-81. Any 2001.

[Seitz-Dyer 95] Steven M. Seitz, Charles R. Dyer, Dept. Computer Sciences, University of Wisconsin, Madison WI, *Phisically-Valid View Synthesis by Image Interpolation*, Proceedings Workshop of Representation of Visual Scenes, Cambridge MA. 1995.

[Seitz-Dyer 96] Steven M. Seitz, Charles R. Dyer, Dept. Computer Sciences, University of Wisconsin, Madison WI, *Toward Image-Based Scene Representation Using View Morphing*, Proceedings Intl. Conference on Pattern Recognition (ICPR'96) Viena 1996.

[Seitz 97] Steven M. Seitz, Dept. Computer Sciences, University of Wisconsin, Madison WI, *Image-Based Transformation of Viewpoint and Scene Appearance*, Tesis Doctoral de S. Seitz, University of Wisconsin. 1997.

[Shannon 48] C.E. Shanon *A mathematical theory of comunication.* Bell System Technology Journal, número 23. pàgines 379-423. 1948

[Slabaugh 01] G. Slabaugh, B. Culbertson, T. Malzbender, and R. Schafer. *A survey of methods for volumetric scene reconstruction from photographs.* Technical Report 1, Center for Signal and Image Processing, Georgia Institute of Technology, 2001.

[Studholme 95] C. Studholme, D. Hill, D. Hawkes, *Multiresolution Voxel Similarity measures for mr-pet registration,* proceedings of the Information Processing in Medical Imaging Conference. 1995.

[Szeliski 93] R. Szeliski. *Rapid octree construction from image sequences.* Computer Vision, Graphics and Image Processing: número 58(1), pàgines 23–32. 1993.

[Szeliski 99] R. Szeliski and P. Golland. *Stereo matching with transparency and matting.* Int. J. of Computer Vision, número 32(1), pàgines 45–61. 1999.

[Watson 99] Andrew B. Watson, James Hu, John F McGowan, Jeffrey B. Mulligan. *Design and performance of a digital video quality metric*. Research Report. NASA Ames Research Center. 1999.

[Wellner 93] Pierre Wellner, *Interacting with paper on the DigitalDesk*, Communications of the ACM, 1993, volum 36, número 7, pàgines 87 a 96.

[Winkler 99] Stefan Winkler, *Issues in Vision Modeling for Perceptual Video Quality Assesment*. Elsevier Preprint, Laboratoire de Traitement des Signaux, EPFL, 1999.

[Wolberg 90] George Wolberg, *Digital Image Warping,* IEEE Computer Society Press. 1990.

[Yang-Welch-Bishop 02] R. Yang, G. Welch, and G. Bishop. *Real-time consensus based scene reconstruction using commodity graphics hardware*. Proceedings of Pacific Graphics. 2002

[Zhu-Riseman 01] Z. Zhu, E.M. Riseman, A.R. Hanson, *Theory and practice in making seamless stereo mosaics from airbone video*. CS-Technical Report, 01-01. Umass-Amherst, Gener 2001.

Part 2. Referències a la xarxa.

Les següents referències corresponen a adreces de la *world wide web* verificades a data de Juny de 2006. Malgrat la volatilitat dels noms o adreces a la xarxa, les referenciades aquí han demostrat tenir una estabilitat suficient com per preveure la seva utilitat en el futur.

[ARQuake 06] http://www.tinmith.net/arquake

[Arvika 03] http://www.arvika.de

[AS 99] http://www.actuality-systems.com/ Actuality Systems Inc. Cambridge MA. *Choosing a True 3-D Display Solution*, 1999.

[ATI 06] http://www.ati.com/products

[BBC 00] http://www.bbc.co.uk/dinosaurs/ BBC: *Walking with dinosaurs*

[Boeing 97] http://www.boeing.com/assocproducts/art/tech_focus.html Boeing: Mathematics and Computing Technology. (Consulta: Juny 2006)

[CompVision 06] http://www.cs.cmu.edu/~cil/vision.html *The Computer Vision HomePage*, suportat per la Carnegie-Mellon University desde 1994, recull dels grups de treball en Computer Vision a nivell mundial.

[Criminisi 05] http://plus.maths.org/issue23/features/criminisi/ Antonio Criminisi and Rachel Thomas, *Getting into the picture*

[Faugueras 99] http://www-sop.inria.fr/robotvis/levelsets/stereo.html

[grec 06] http://www.grec.net *Gran diccionari de la llengua catalana en línia*. Grup Enciclopèdia Catalana.

[iec 06] http://www.iec.cat *Diccionari normatiu*. Institut d'Estudis Catalans.

[Kanade 99] http://www.cs.cmu.edu/~virtualized~reality/ *Virtualized Reality Home Page*, Carnegie Mellon University

[Laveau 99] http://www-sop.inria.fr/robotvis/personnel/laveau Stephane Laveau: *Generating Views Without 3-D Models*

[McMillan 99] http://www.cs.unc.edu/~mcmillan/ Leonard McMillan, Dept. Computer Sciences, University of North Carolina

[mpeg 06] http://bmrc.berkeley.edu/frame/research/mpeg/faq/mpeggeneral.html

[NVidia 06] http://www.nvidia.com/page/workstation.html

[Polhemus 05] http://www.polhemus.com

[SG 06] http://www.sgi.com/products

[Seitz 99] http://www.cs.wisc.edu/~seitz/interp/interp.html

[termcat 06] http://www.termcat.cat Centre català de terminologia.

[XBox 06] http://www.xbox.com/en-US/hardware