Universitat de les Illes Balears

Departament de Ciències Matemàtiques i Informàtica

Tesi Doctoral

Reconeixement visual del moviment humà en temps real per a la interacció natural home-màquina

Antoni Jaume-i-Capó

Dirigida per:

Dr. Francisco Perales López

Dr. Javier Varona Gómez

18 de juny 2009

Dr. Francisco José Perales Lopez.

Professor Titular d'Universitat.

Departament de Ciències Matemàtiques i Informàtica.

Universitat de les Illes Balears.

Dr. Javier Varona Gómez.

Ramón y Cajal.

Departament de Ciències Matemàtiques i Informàtica.

Universitat de les Illes Balears.

FAN CONSTAR:

Que la memòria titulada *Reconeixement visual del moviment humà en temps real per a la interacció natural home-màquina* ha estat realitzada per Antoni Jaume i Capó baix la nostra direcció en el Departament de Ciències Matemàtiques i Informàtica de la Universitat de les Illes Balears i constitueix la tesi per optar al grau de Doctor en Informàtica.

Palma, 18 de juny de 2009

Dr. Francisco José Perales López

Director de la tesi

Dr. Javier Varona Gómez

Director de la tesi

Antoni Jaume i Capó

Doctorant

A totes i a tots.

Agraïments

Acabada la memòria de la tesi, m'agradaria mostrar el meu agraïment cap a les següents persones:

- A la meva família, que m'ha permès el luxe de seguir aquest camí, i els quals m'han ensenyat molt. En especial als meus pares i als meus germans.
- A na Margalida, perquè m'ajuda a desconnectar.
- A tots els amics, per *lo* mal de sofrir que sóc i *lo* mal de sofrir que són. Segur que quan siguem majors contarem al jovent les nostres batalles de joventut. Perquè no perdem mai aquest esperit.
- A tots els companys i amics del Laboratori, de la Unitat, del Departament i de la Universitat, que els sobra la feina però sempre tenen un moment per ajudar.
- I molt especialment als directors, pels ànims i els consells.

Sempre hi ha coses per fer.

Abstract

In most of the existing human-computer interfaces, enactive knowledge as new natural interaction paradigm has not been fully exploited yet. Recent technological advances have created the possibility to enhance naturally and significantly the interface perception by means of visual inputs, the so-called Vision-Based Interfaces (VBI).

In the present document, first, we explore the recovery of the user's body posture by means of combining robust computer vision techniques and a well known inverse kinematics algorithm in real-time. The 3D position of the hands are extracted in real-time and provided to the body posture recovery algorithmic layer. This motion capture system is capable to estimate the user 3D body joints position in real-time. We focus the tests in terms of performances and overall quality of the reconstructed body posture

Then, we present a gesture recognition algorithm where the user's movements are obtained through the real-time vision-based motion capture system. Specifically, we focus on recognizing users motions with a particular mean, that is, a gesture. Defining an appropriate representation of the user's motions based on a temporal

posture parameterization, we apply non-parametric techniques to learn and recognize the user's gestures in real-time. This scheme of recognition has been tested for controlling a classical computer videogame. The results obtained show an excellent performance in on-line classification and it allows the possibility to achieve a learning phase in real-time due to its computational simplicity.

Finally, we present how to add image constraints to inverse kinematics in order to improve the results of the real-time vision-based motion capture system. Specifically, we explain how to define a criterion to use images in order to guide the posture reconstruction of the articulated chain. Tests with synthetic images show how the scheme performs well in an ideal situation. In order to test its potential in real situations, more experiments with real images are also presented. By means of a quantitative study of different sequences, the results obtained show how this approach improves the performance of inverse kinematics in this application.

Key words: Enactive interfaces; Human-computer interaction; Vision-based interfaces; Vision-based gesture recognition; Inverse kinematics; 3D reconstruction from images.

Resum

En la majoria d'interfícies persona-ordinador existents, el coneixement *Enactiu* com a nou paradigma d'interacció natural, a dia d'avui, no ha estat completament aprofitat. Novetats tecnològiques recents han creat la possibilitat de millorar naturalment i significa, la percepció de la interfície a través d'entrades visuals, les anomenades interfícies basades en visió (VBI).

En aquest treball, en primer lloc, s'explora la recuperació de la postura del cos de l'usuari utilitzant la combinació de tècniques robustes de visió per ordinador i els ben coneguts algorismes de cinemàtica inversa en temps real. Aquest sistema de captura del moviment es capaç d'estimar les posicions 3D de les articulacions de l'usuari en temps real. Els experiments es centren en demostrar la qualitat global de la postura recuperada.

Llavors, es presenta un algorisme de reconeixement de gestos on els moviments de l'usuari s'obtenen a través del sistema de captura del moviment en temps real basat en visió. Concretament, aquesta part del treball es centra en reconèixer els moviments de l'usuari que tenen un sentit particular, o sigui un gest. Definint una representació apropiada dels moviments de l'usuari, basada en una parametrització temporal de la

postura, s'apliquen tècniques no paramètriques per aprendre i reconèixer els gestos de l'usuari en temps real. L'esquema s'ha provat per controlar un videojoc. Els resultats obtinguts mostren un excel·lent funcionament de la classificació en temps real. A més, permeten realitzar una fase d'aprenentatge en temps real, a causa de la seva simplicitat computacional.

Finalment, es presenta com afegir una restricció basada en la imatge a la cinemàtica inversa, amb l'objectiu de millorar els resultats del sistema de captura en temps real basat en visió. Concretament, s'explica com definir un criteri per utilitzar imatges, amb la finalitat de guiar la reconstrucció de la postura de la cadena cinemàtica. Proves amb imatges sintètiques mostren que la proposta funciona correctament en una situació ideal. Amb l'objectiu de provar el seu potencial en situacions reals, es presenten més experiments amb imatges reals. Mitjançant un estudi quantitatiu de diferents seqüències, els resultats obtinguts mostren que l'enfocament millora el rendiment de la cinemàtica inversa en aquesta aplicació.

Paraules clau: Interfícies enactives; Interacció persona-ordinador; Interfícies basades en visió; Reconeixement de gestos basat en visió; Cinemàtica inversa; Reconstrucció 3D a partir d'imatges.

Índex

Índex de figures

Índex de taules

Capítol 1

Introducció

Tot comença ...

Un reconegut psicòleg va proposar a uns nins un joc que consistia en treure un objecte d'una caixa transparent sense apropar-se a ella. Per fer-ho tenien que utilitzar bastons, cordes i pinces. Va organitzar els nins en tres grups. Als del primer grup els va deixar jugar lliurament amb els objectes abans d'explicar-los en què consistia el joc. Als del segon grup, no els va deixar jugar amb el material però els va fer una demostració a l'hora d'explicar-los en que consistia el joc. Finalment, als del tercer grup els va deixar jugar un poc i els va donar unes quantes explicacions. El resultat va ser molt il·lustratiu, els nins del primer grup foren els que abans i millor acabaren el joc. Aquest experiment té moltes interpretacions possibles, però la que és més interessant és l'adquisició de coneixement per experimentació, i un exemple clarificador el podem trobar en l'ensenyança d'esports, ja que per molt que

li expliquin a un nin com ha de jugar a futbol o bàsquet, com realment aprèn més és practicant l'esport. Aquest enfocament es defineix com coneixement *Enactiu* [70].

El coneixement *Enactiu* representa un tipus de coneixement per experimentació, basat en les respostes perceptuals a les accions de l'usuari, adquirit i perfeccionat a través de la pràctica. El terme *Enactiu* prové etimològicament de l'anglès *to enact* (representar), i no té un terme equivalent per traduir-lo, per tant, enactuar significaria la possibilitat de presentar i actuar al mateix temps, en el nostre cas, es podria interpretar com *veure i fer a la vegada*.

Encara que fins ara les tecnologies d'interacció persona-màquina no han utilitzat completament aquest potencial del coneixement *Enactiu*, avanços tecnològics recents permeten la possibilitat d'enriquir de forma natural la percepció de la interfície utilitzant entrades visuals. Definides com interfícies basades en visió (VBI, de l'anglès Vision-based interfaces), proposades per Turk et al. [68], utilitzen tècniques de visió per ordinador amb l'objectiu de detectar i percebre l'usuari i les seves accions en un context d'interacció persona-ordinador (HCI, de l'anglès Human Computer Interaction).

La visió per ordinador és la disciplina científica que intenta que els ordinadors percebin la informació visual a través de l'anàlisi d'imatges o seqüències [25]. A dia d'avui, la tecnologia de la visió per ordinador aplicada a interfícies persona-ordinador està tenint un èxit important [43]. L'avantatge d'aquestes interfícies és que els dispositius d'adquisició són passius i no intrusius, és a dir, no requereixen contacte amb l'usuari. Des del punt de vista de la interacció persona-ordinador, l'interès es centra en obtenir els moviments de l'usuari, per posteriorment reconèixer

els que poden ser interpretats com esdeveniments del sistema.

De fet, la informació visual és molt important quan les persones interaccionen entre elles i el seu entorn. Mitjançant la visió, les persones són capaces de determinar la localització, la identitat, l'estat d'ànim, l'activitat o fins i tot aproximar l'edat de l'interlocutor. Aquestes entrades visuals poden afectar el contingut i el flux de la conversació, proporcionant informació contextual de forma diferent a la parla, inclús permeten evitar ambigüitats com per exemple és el cas de la paraula *això* mentre es senyala un objecte de l'entorn. Per altra banda, si es parla de comunicació en general, com el procés que permet a les persones que intercanviïn informació, els experts la classifiquen en dues formes: la verbal i la no-verbal [35]. La comunicació no-verbal és entesa com el procés d'enviar i rebre missatges no orals.

S'ha demostrat que la comunicació no-verbal és molt més ample que la verbal, i que moltes vegades s'utilitzen les dues al mateix temps. És un fet, que les persones en tot moment emeten informació no-verbal mitjançant els gestos, les postures, les expressions facials, la manera de vestir, el pentinat, l'entonació, ... La majoria d'aquesta comunicació no-verbal és visual, i una de les formes més importants són els gestos, que es porten a terme de forma conscient amb el cos, i és una forma molt rica que permet als individus expressar una gran varietat de sentiments i pensaments, veure Figura 1.1.

El problema a l'hora de reconèixer el moviment corporal d'una persona és la variabilitat en què diferents persones realitzen diferents moviments que posseeixen el mateix significat. A més, des d'un punt de vista computacional s'afegeix la dimensió temporal i la complexitat del número de graus de llibertat del moviment del cos

Figura 1.1: Gest per demanar silenci.

humà. En aquest cas, l'aproximació més utilitzada per resoldre aquest problema és fer el reconeixement depenent de l'aplicació. Un exemple il·lustratiu, es pot trobar en els algorismes de captura de moviment humà on es limiten els moviment que pot fer un usuari a un conjunt d'accions apreses prèviament [15, 69]. Una altra aproximació possible, és utilitzar la visió per ordinador per recuperar el moviment de només algunes parts del cos, en el cas que la part del moviment posseeixi informació suficient per a la interfície, i a més es redueix notablement la complexitat del problema [50]. L'exemple més clar, es troba en el reconeixement de signes de l'alfabet per a sords [49]. De totes formes, per reconèixer accions o gestos més complexes és necessari recuperar

la postura de tot el cos.

Concretament, les propostes existents pel reconeixement de gestos basat en visió, poden ser classificades en tres grans categories: les basades en moviment, les basades en aparença i les basades en model. Les propostes basades en moviment intenten reconèixer el gest directament des del moviment sense utilitzar cap informació estructural sobre el cos físic [53, 4, 23]. Les propostes basades en aparença usen informació bidimensional tal com imatges en escala de grisos, contorns o siluetes del cos [64, 24]. Finalment, les propostes basades en model es centren en recuperar la configuració tridimensional de les parts del cos articulat [57, 72, 59, 36].

És evident que si es pot recuperar la postura de l'usuari, és la manera més senzilla de poder analitzar i interpretar els seus moviments, ja que es disposaria de la descripció completa dels moviments 3D de l'usuari. Però el problema dels enfocament basats en el model, és que moltes vegades són difícils d'aplicar al món real. Principalment, per la dificultat de capturar i seguir totes les parts necessàries del model, o sigui les articulacions de l'usuari que participen en els gestos. A més, si la finalitat és la interacció, els algorismes han de treballar en temps real i la majoria dels que existeixen no hi treballen.

En aquest sentit, hi ha treballs que han demostrat que és possible recuperar la postura completa d'un usuari en temps real a partir de parts terminals del cos (anomenades *end-effectors*), com poden ser per exemple les mans [8]. Per això, en aquest treball es presenta un enfocament de reconeixement de gestos basat en el model, on en primer lloc es recupera la postura de l'usuari en temps real i a continuació es reconeixen els gestos que l'usuari realitza.

Per obtenir els moviments de l'usuari, recuperant la seva postura, es presenta un enfocament basat en visió que combina l'anàlisi de les imatges proporcionades per dues càmeres (observació) i un algorisme de cinemàtica inversa (control). Analitzant les imatges d'entrada es realitza el seguiment visual 3D dels *end-effectors*, en temps real. A continuació, amb l'algorisme de cinemàtica inversa i a partir de la posició 3D dels *end-effectors*, s'estimen el resta d'articulacions del cos de l'usuari.

A partir de la posició 3D de les articulacions de l'usuari, es presenta un sistema que és capaç de reconèixer un conjunt de gestos, solucionant els principals problemes en el desafiament del reconeixement de gestos: les variacions temporals, les variacions espacials i les variacions d'estil. Les variacions temporals són causades per la diferència de velocitat a l'hora de realitzar els gestos entre diferents usuaris. Les variacions espacials són degudes a les restriccions físiques del cos humà, com són les diferents talles. Les variacions d'estil són causades per la forma personal en què els usuaris realitzen els seus moviments. Aquest darrer és el repte més important, ja que el sistema que es presenta permet a l'usuari treballar amb els gestos que ell consideri naturals per cada esdeveniment d'interacció, i no se l'obliga a utilitzar un conjunt de gestos predefinits. El fet d'utilitzar gestos naturals, està molt relacionat amb el coneixement *Enactiu* basat en l'experimentació, on l'usuari per cada esdeveniment ha d'utilitzar els gestos més lògics a partir de la seva experiència.

1.1 Objectius

L'objectiu principal d'aquesta memòria és estudiar i realitzar una interfície persona-ordinador, que permeti a l'usuari interaccionar mitjançant gestos naturals a través del reconeixement visual del seu moviment. Per això, les dues parts més importants d'aquest treball són:

- Obtenir la postura completa 3D de l'usuari en temps real, utilitzant tècniques de visió per computador i cinemàtica inversa.

- Realitzar un algorisme que permeti reconèixer els moviments de l'usuari que corresponen a gestos, a partir de la postura 3D, per a la interacció natural persona-ordinador.

1.2 Organització de la memòria

En el primer capítol es presenten les idees principals i objectius d'aquest treball. En el Capítol 2, *Sistema de captura dels moviments de l'usuari*, es descriu com es capturen els moviments de l'usuari amb l'objectiu de reconèixer els seus gestos.

A continuació, en el Capítol 3, *Reconeixement de gestos per a la interacció natural*, es descriu com es reconeixen els gestos de l'usuari a partir de les postures capturades.

En el Capítol 4, *Restricció basada en la imatge per a la cinemàtica inversa*, es presenta una nova restricció basada en la imatge per millorar els resultats del sistema captura del moviment humà.

Finalment, en el Capítol 5, *Conclusions*, s'enumeren les conclusions que es poden

extreure d'aquest treball, idees de futur i un llistat d'articles i projectes del qual aquest treball ha format part.

Endemés al final del document, es presenta la bibliografia utilitzada i un Apèndix sobre el modelat automàtic del cos de l'usuari.

Capítol 2

Sistema de captura dels moviments de l'usuari

Els homes no viven junts perquè si,
sinó per dur a terme grans empreses.
José Ortega y Gasset.

En aquest capítol es descriu com es capturen els moviments de l'usuari, que més endavant s'utilitzaran per reconèixer els seus gestos. L'objectiu principal és obtenir les posicions 3D de les seves articulacions, per aquest motiu es modela el seu cos mitjançant una cadena cinemàtica, que consisteix en un conjunt de segments units per articulacions. Per al control de la cadena cinemàtica s'utilitza la cinemàtica inversa (IK), que permet estimar la seva configuració a partir d'algunes posicions conegudes, anomenades *end-effectors*, que s'obtenen utilitzant algorismes de visió

mitjançant la captura de l'usuari a través d'un parell estèreo de càmeres.

2.1 Enfocament

Per poder reconèixer els gestos de l'usuari és necessari reconstruir la seva postura, ja que el sistema de reconeixement de gestos que es presenta en aquest treball es basa en el model, com s'ha explicat al Capítol 1. Endemés, perquè l'usuari pugui realitzar gestos naturals és necessari que el sistema de captura no sigui invasiu.

Fins fa poc, l'ús de la captura en temps real del moviment del cos de l'usuari estava limitada a aplicacions com l'animació expressiva d'un personatge virtual [65]. Entre altres factors, perquè l'adopció dels moviments del cos humà com interfície d'usuari 3D ha estat obstaculitzada per els següents motius, entre d'altres: s'utilitzen sensors invasius, l'espai d'adquisició és molt limitat, la distorsió espacial i la gran dimensió de l'espai de la postura. Aquests factors són font d'error i s'acumulen provocant com a resultat una postura aproximada, que en el cas de l'animació és suficient, però si es desitja un control espacial no és adequat per a una interacció complexa.

En aquest treball es presenta un sistema de captura del moviment de l'usuari no-invasiu basat en visió, que recupera la postura de l'usuari estimant la posició de les seves articulacions. Aquesta captura de moviments no es pot realitzar només amb tècniques de visió, ja que les imatges capturades poden tenir renous o ser incompletes. Per una part les extremitats i/o articulacions de l'usuari poden estar auto-ocluïdes, ja que l'usuari pot situar una mà darrera el seu tronc o pot posar una mà sobre l'altra. Per altra banda, la roba de l'usuari pot dificultar saber amb claredat on està

situada una articulació. Per aquests motius el sistema de captura dels moviments de l'usuari combina algorismes de visió per ordinador i de cinemàtica inversa, veure Figura 2.1.

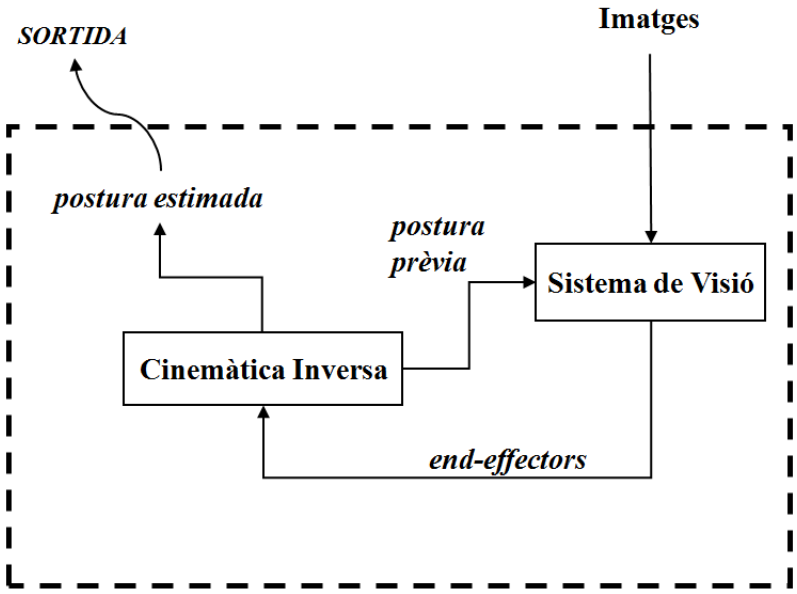

Figura 2.1: Arquitectura general del sistema.

A l'hora de reconstruir la postura de l'usuari és habitual modelar-la utilitzant una cadena cinemàtica [9, 8], que com s'ha explicat anteriorment consisteix en un conjunt de segments rígids, units per articulacions. La cinemàtica inversa permet estimar la configuració de la cadena cinemàtica, que modela l'usuari, a partir de posicions conegudes que s'anomenen *end-effectors*. Aquests *end-effectors* són obtinguts a partir

de dues càmeres estàndard que capturen imatges en color , que mitjançant algorismes de visió per ordinador els localitzen en temps real, en el cas d'aquest treball són les seves mans. Tot seguit, l'algorisme de cinemàtica inversa utilitza aquests *end-effectors* per estimar la postura de l'usuari, mitjançant la cadena, per cada parell d'imatges capturades.

2.2 Treballs previs

La captura en temps real del moviment del cos humà té una llarga història a l'hora de realitzar animacions [65]. La manca de facilitat d'ús per part de la tecnologia de l'exosquelet ha evitat que s'estengués àmpliament. En la dècada dels anys 90 va aparèixer la tecnologia dels sensors magnètics, veure Figura 2.2 gràcies al treball de Badler, on utilitzava quatre sensors magnètics (a la cintura, al cap i a les dues mans) per conduir la postura d'un model humà amb cinemàtica inversa [1]. L'objectiu era recrear la postura humana amb els mínims sensors possibles sobre l'usuari. Així i tot, els graus de llibertat no controlats com l'angle de gir dels braços, amb el temps provoquen diferències importants entre l'usuari real i el seu model virtual. Més endavant, Molet va descriure un proposta per eliminar aquesta ambigüitat utilitzant més sensors [45] per intentar minimitzar els errors que provocaven els graus de llibertat no controlats. Existeixen enfocaments similars, on s'identifica l'estructura de l'esquelet i les longituds dels segments [5, 47].

Treballs més recents mostren un renovat interès en proposar enfocaments menys invasius que utilitzin un reduït nombre de sensors [27, 15]. Aquests enfocaments,

primer aprenen els models locals del moviment humà a partir d'una base de dades de captura de moviments predefinits. Després, usant aquests models locals, a partir de les imatges d'entrada capturades utilitzant marcadors retro-reflectius, es guia la recuperació del moviment, veure Figura 2.3. Cal destacar, que aquestes tècniques no s'usen per la interacció 3D, bàsicament perquè els moviments de l'usuari estan restringits als moviments prèviament apresos. Una altra possibilitat és compensar la falta d'informació a través de restriccions. La possibilitat d'associar prioritats estrictes a les restriccions és l'aspecte clau per tenir èxit, com es destaca en el context de l'optimització interactiva de la postura [2]. Alternativament, l'enfocament analític de la cinemàtica inversa és generalment més eficient en termes de cost de computació però no permet assignar nivells de prioritat a les restriccions [66].

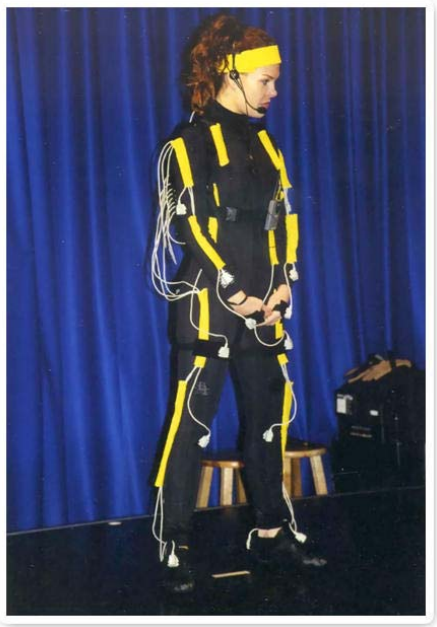

Figura 2.2: Sistema de captura del moviment humà que utilitza sensors magnètics.

La reconstrucció del moviment d'una persona a partir de l'analisis d'imatges ha rebut gran atenció en els darrers anys en el camp de la visió per ordinador [44, 72, 43]. No obstant, la majoria de les propostes actuals no treballen en temps real, per tant dificulten la comparació amb les que ho son. Per aquest treball, el temps real és una restricció molt importat, ja que l'objectiu és utilitzar les postures capturades com entrada de la interfície d'usuari per a la interacció persona-màquina. Un treball interessant que *a priori* treballa en temps real és el de Wren [74] del Medialab del M.I.T. En aquest treball, els autors presenten un sistema de seguiment 3D de la part superior del cos humà situat enfront d'un dispositiu de realitat virtual. Així i tot, no es presenta cap avaluació del sistema. Endemés els possibles gestos es restringeixen a un conjunt predefinit de moviments apresos prèviament. Aquesta aproximació redueix l'espai de cerca dels moviment humans.

Figura 2.3: Sistema de captura del moviment humà que utilitza marcadors retro-reflectius.

2.3 El sistema de visió

A partir d'un parell estèreo de càmeres es capturen el moviments de l'usuari en l'espai de captura. Per poder estimar la postura de l'usuari per cada parell d'imatges capturades, és important localitzar correctament els *end-effectors*, que en aquesta proposta són les mans de l'usuari. Per aconseguir-ho, s'utilitzen algorismes d'eliminació de fons, de segmentació per color de pell i de seguiment 2D d'ambdues mans de l'usuari en cada imatge. Llavors, es combina aquest resultat amb un algorisme de seguiment 3D per estimar de forma robusta les posicions 3D dels *end-effectors* de l'usuari en l'escena.

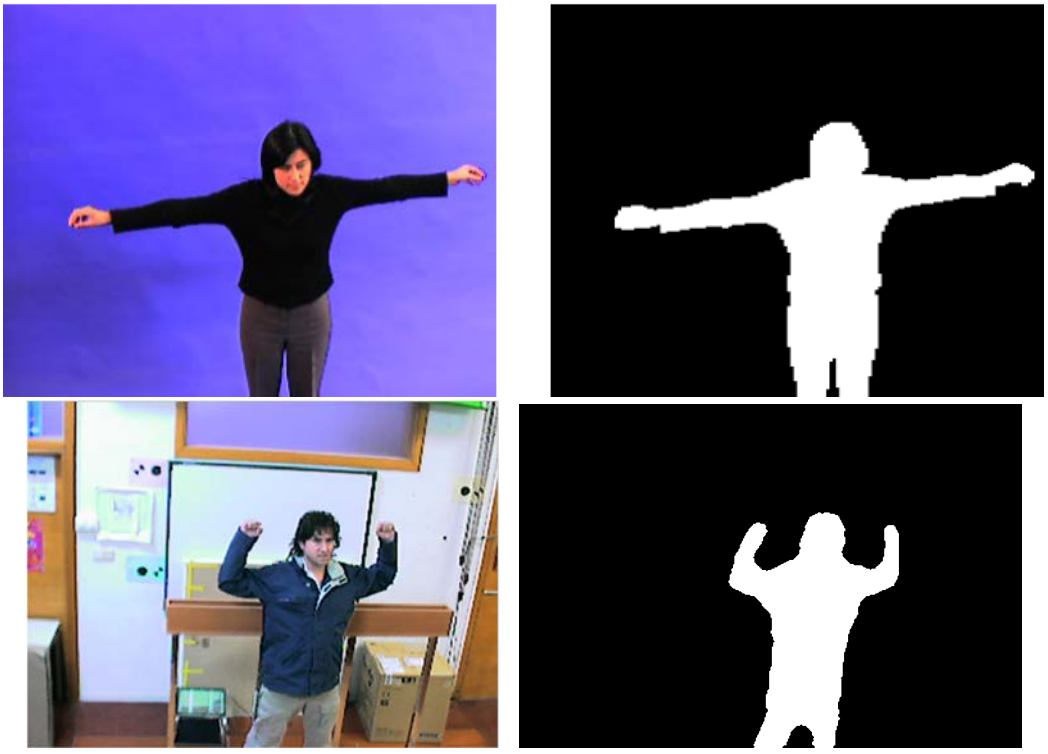

Figura 2.4: Eliminació de fons.

En primer lloc i per cada imatge del parell estèreo, amb l'objectiu de conèixer els píxels que formen part de l'usuari i els que no, s'apliquen algorismes d'eliminació de fons [63, 32] sobre les imatges originals. Aquests algorismes construeixen una màscara amb els píxels de la imatge que pertanyen a l'usuari, veure la Figura 2.4.

Una vegada es saben quins píxels de la imatge són de l'usuari, s'ha de detectar quins d'aquests píxels corresponen a les mans, ja que són les posicions que s'utilitzen com *end-effectors*. Per aconseguir-ho, s'utilitza el color de la pell de l'usuari per segmentar les mans i la cara [11, 19]. Aquest algorisme de segmentació necessita una fase prèvia d'aprenentatge del model de color de pell de l'usuari, que habitualment consisteix en una inicialització manual on es seleccionen les regions de l'usuari que contenen píxels de color de pell per crear una mostra [71].

$$\{R > 95 \text{ AND } G > 40 \text{ AND } B > 20 \text{ AND } \max(R,G,B) \cdot \min(R,G,B) > 15 \text{ AND } |R \cdot G| > 15 \text{ AND } R > G \text{ AND } R > B\}$$
$$\text{OR}$$
$$\{R > 220 \text{ AND } G > 210 \text{ AND } B > 170 \text{ AND } |R \cdot G| \leq 15 \text{ AND } R > B \text{ AND } G > B\}$$

Taula 2.1: Normes heurístiques.

Perquè la selecció de la mostra es faci de forma automàtica, en aquest treball es proposa utilitzar l'algorisme [37] que detecta el píxels de color de pell de forma automàtica. Aquest algorisme utilitza un espai de color 3D (RGB), que amb l'ajuda de normes heurístiques determina si un píxel de la imatge correspon al color de la pell (les normes heurístiques es mostren a la Taula 2.1). Amb aquestes normes es pot assegurar que tots els píxels detectats són de color de pell, però pot passar que píxels de color de pell no siguin detectats. Per aquest motiu, en el primer parell d'imatges

capturades d'una sessió d'un usuari, s'utilitzen els píxels detectats per les normes heurístiques com a llavor per agafar la mostra de píxels de color de pell. Aquests són agrupats en el·lipses i tots els píxels continguts dins les el·lipses es consideren la mostra del color de pell de l'usuari.

Una vegada que tenim la mostra de píxels de color de pell, tant si s'ha obtinguda automàticament com manual, es transformen a l'espai HSL, per poder treballar amb el tò i la saturació, o sigui la cromacitat. El valors de la mostra de píxels són utilitzats per construir el model de color de pell:

$$\mathbf{X} = (x_1, \ldots, x_n) \tag{2.1}$$

on n és el nombre de mostres i $\boldsymbol{x_i} = (h_i, s_i)$, on h és el tò i s la saturació. Com a model estadístic s'utilitzà el gaussià, ja que s'ha demostrat dóna bons resultats [71]:

$$\boldsymbol{\mu} = \frac{1}{n}\sum_i \boldsymbol{x_i},\ \boldsymbol{\sigma^2} = \frac{1}{n}\sum_i (\boldsymbol{x_i} - \mu_i).(\boldsymbol{x_i} - \mu_i)^T \tag{2.2}$$

A partir del model estadístic del color de pell, es pot calcular la probabilitat que un píxel sigui de color de pell:

$$P(x) = \frac{1}{\sqrt{(2\pi)^2|\sigma^2|}} e^{\frac{1}{2}(x-\mu)(x-\mu)^T} \tag{2.3}$$

on $|\cdot|$ és el determinant de la matriu corresponent.

Amb el model estadístic dels píxels de color de pell i la seva funció de probabilitat, per cada parell estèreo capturat, es calcula la probabilitat per tots els píxels de la

imatge per detectar quins píxels són de color de pell. A aquests píxels s'aplica un algorisme de connexió de components per agrupar-los en blobs, veure Figura 2.5.

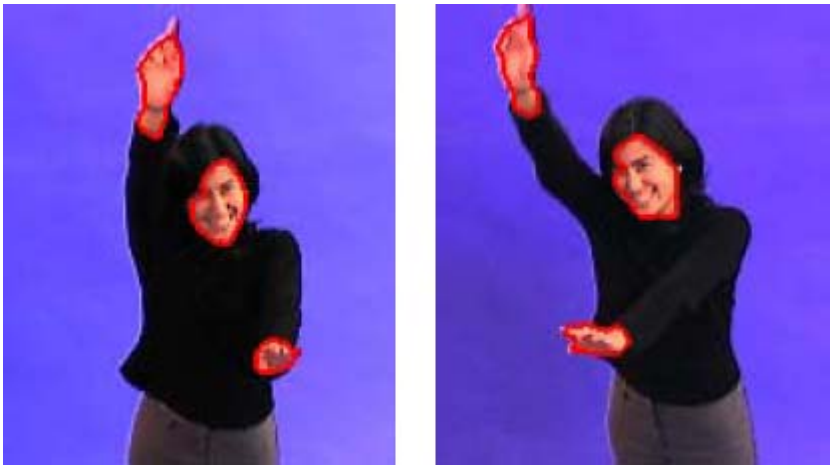

Figura 2.5: Segmentació dels blobs color de pell de l'usuari.

La següent passa del sistema de visió és conèixer a quina part del cos correspon cada blob de color de pell en cada imatge. Per aquest motiu s'utilitza un algorisme que etiqueta els blobs a partir d'un conjunt d'hipòtesis d'imatges anteriors [71]. Per una imatge de l'instant t i amb les etiquetes de l'instant de temps $t-1$ s'etiqueten quins blobs de la imatge corresponen a la cara, la mà dreta i la mà esquera; a més es detecta si un blob de color de pell entra en l'espai de captura o desapareix. En definitiva, a partir de les hipòtesis de l'instant anterior s'etiqueten els blobs dels píxels de color de pell de la imatge actual. La Figura 2.6 mostra els resultats finals d'aquest procés.

La darrera fase per localitzar la posició de les mans i la cara, una vegada s'han localitzat en cada imatge del parell estèreo, és calcular la seva posició 3D. Per fer-

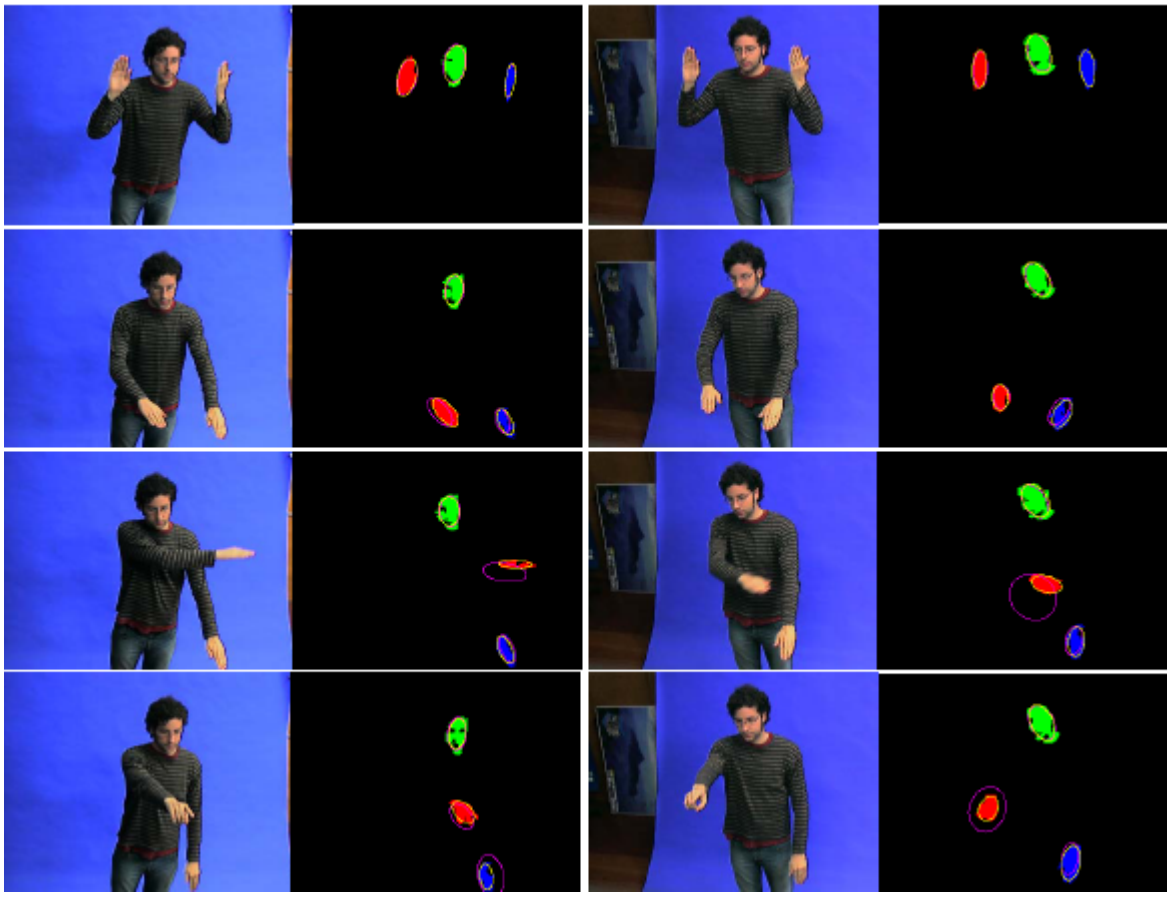

Figura 2.6: Blobs etiquetats.

ho, es calcula usant el mètode de triangulació del punt mig, però abans és necessari conèixer els paràmetres intrínsecs i extrínsecs del conjunt de càmeres, per això, abans que el sistema comenci a funcionar s'ha de calibrar. Per fer-ho, en aquest treball s'utilitza un patró pla, veure Figura 2.7, per calcular els paràmetres intrínsecs i extrínsecs del parell de càmeres estèreo [77].

El paràmetres intrínsecs fan referència al model de la càmera: distància focal, punt principal (centre del sensor de la càmera), coeficients de biaix (angle que for-

Figura 2.7: Patró pla de calibratge.

men els eixos x i y del sensor), i distorsions (coeficients radials i tangencials de distorsió [29]). Aquests paràmetres no canvien si la posició o orientació de la càmera varia. Per tant mentre es treballi amb el mateix model de càmeres i òptica només serà necessari calcular-los una vegada.

Per altra banda, els paràmetres extrínsecs, fan referència a la posició i la orientació de les càmeres en un espai 3D, amb un mateix sistema de referència. Per aquest motiu sempre que es variï la posició d'alguna càmera del sistema, és necessari tornar a calcular aquests paràmetres.

Coneixent els paràmetres de calibratge, es calcula la posició 3D dels blobs projectant la posició 2D de cada blob en cada imatge a l'infinit, i teòricament la intersecció d'aquestes línies és la posició 3D que es cerca. La problemàtica que presenta aquest esquema de triangulació, és que en el procés de localització dels blobs i en el procés de calibratge hi sol haver alguns petits errors, que provoquen que aquestes línies poques vegades interseccionin. Llavors, aquesta posició no es pot calcular sinó que

s'ha d'estimar.

Existeixen molts de mètodes per trobar aquesta estimació, com es mostren al treball den Hartley [28]. Però per a la nostra proposta és suficient usar el mètode del punt mig [67]. El mètode del punt mig es defineix de la següent forma. Sigui $\mathbf{O}_1 + t_1\overrightarrow{\mathbf{O}_1\mathbf{p}_1}$, amb $t_1 \in R$ la línia r, que passa a través del punt $\mathbf{O}_1$ que és l'origen de la càmera 1 i $\mathbf{p}_1$ que es la posició d'un píxel en la imatge de la càmera 1. Per altra banda sigui $\mathbf{O}_2 + t_2\overrightarrow{\mathbf{O}_2\mathbf{p}_2}$, amb $t_2 \in R$ la línia s, que passa a través del punt $\mathbf{O}_2$ que és l'origen de la càmera 2 i $\mathbf{p}_2$ que es la posició d'un píxel en la imatge de la càmera 2. Ambdues línies estan expressades en la referència del món de la imatge. Sigui $\vec{w}$ un vector ortogonal a r i s. Aleshores, el problema es redueix a determinar el punt mig P', del segment paral·lel a $\vec{w}$ que uneix r i s, veure Figura 2.8.

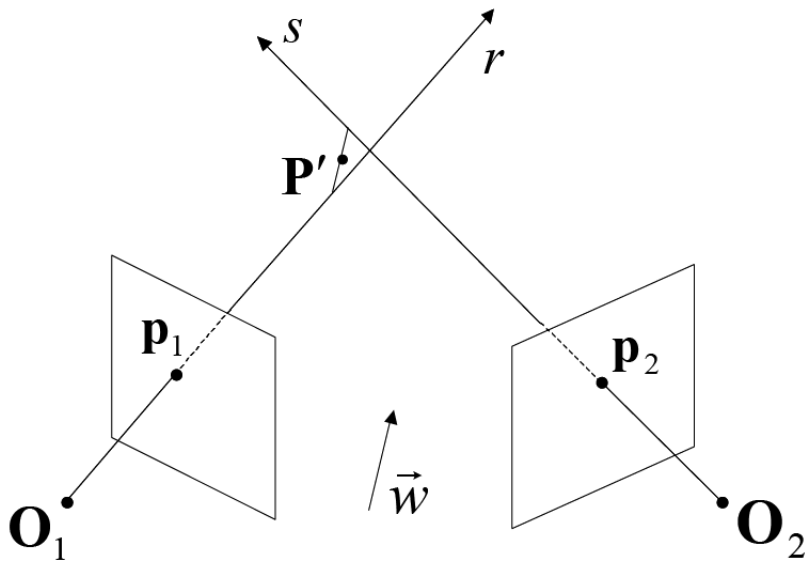

Figura 2.8: Triangulació utilitzant el mètode del punt mig.

També s'aplica un filtre de Kalman [73] per estimar la posició 3D dels blobs a partir de les mesures 2D obtingudes de les imatges. Utilitzar el filtre de Kalman assegura una estimació robusta dels blobs i suavitza les estimacions entre imatges consecutives minimitzant el tremolor de la posició (habitualment referit com a jitter), que poden causar les oscil·lacions sobre l'estimació completa de la postura. També s'usen les prediccions del filtre de Kalman per establir la relació correcte entre cada posició 2D dels *end-effectors*. Per això, primer es triangulen totes les possibles combinacions de les mesures 3D del conjunt d'imatges per obtenir les posicions 3D candidates de cada *end-effector*. Després, per cada *end-effector*, es selecciona el candidat més proper al predit pel filtre d'estimació. La Figura 2.9 mostra els resultats d'aquest procés, retroprojectant la posició 3D de l'*end-effector* associat correctament en les imatges, després d'oclusions severes.

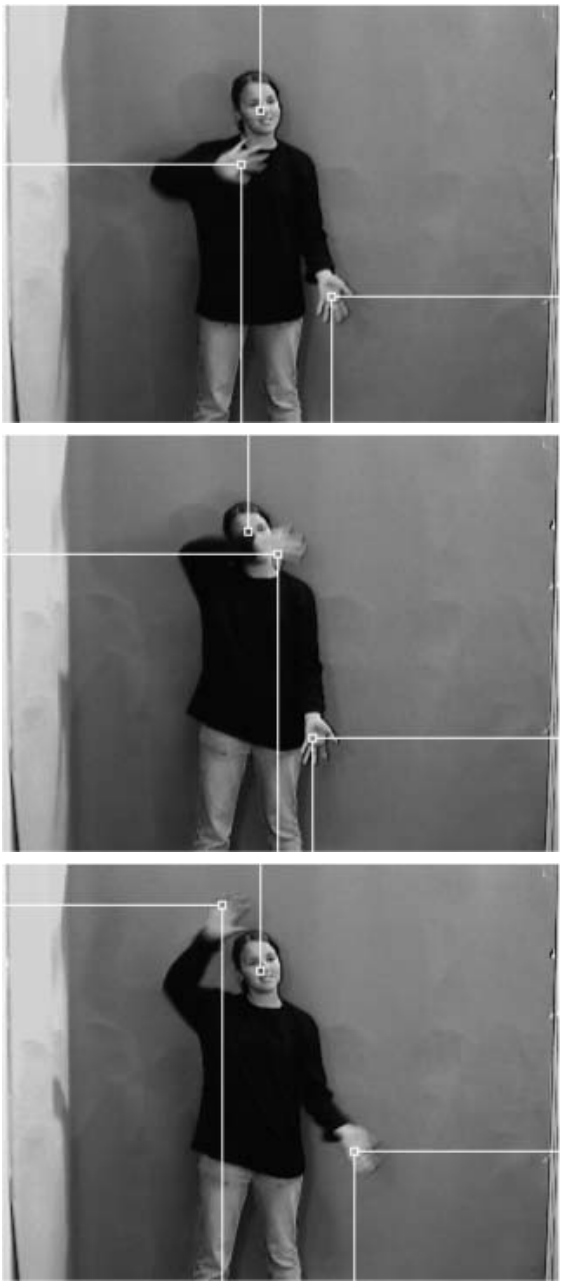

Figura 2.9: Seguiment correcte de les posicions 3D dels *end-effectors* retroprojectats en ambdues imatges. La línia blanca que comença en el límit dret de la imatge correspon a la mà dreta. La línia blanca que comença en el límit esquerra de la imatge correspon a la mà esquerra.

En el cas d'oclusions severes, els blobs no concordaran en ambdues imatges i el resultat de la reconstrucció del punt 3D serà incorrecte. Per aquest motiu i per relacionar robustament els punts 3D a les mesures del conjunt d'imatges es realitza una fase computacional extra. Com que les posicions dels *end-effectors* són en el món 3D, es pot utilitzar un model físic per fer-ne el seguiment i solucionar aquest problema. Un segment en el moment t es caracteritza per la seva posició, que és representada per el vector d'estat $\mathbf{x}_t$. El sistema observa la posició 3D projectada del segment en el vector $\mathbf{z}_t$ (i.e. la posició triangulada de les vistes). La dinàmica del segment és descrita per l'equació en diferències:

$$\mathbf{x}_t = \mathbf{f}_{t,t-1}(\mathbf{x}_{t-1}) + \mathbf{w}_t, \tag{2.4}$$

on $\mathbf{f}_{t,t-1}(\cdot)$ és un vector de la funció que descriu la transició del vector d'estat des de $t-1$ a t, i $\mathbf{w}$ representa l'error del model. La funció de transició d'estat per a un segment és un model polinomial cinemàtic que assumeix una velocitat constant. L'equació de mesura descriu la relació entre la posició observada i les variables d'estat del sistema dinàmic:

$$\mathbf{z}_t = \mathbf{m}_{t,t-1}(\mathbf{x}_t) + \mathbf{n}_t, \tag{2.5}$$

on $\mathbf{m}_{t,t-1}(\cdot)$ és la funció de mesura i $\mathbf{n}$ és el renou de la mesura. Les equacions del filtre de Kalman permeten calcular l'estimació òptima del vector de l'estat recursivament a partir de les mesures i l'estimació inicial. Amb aquest objectiu, en primer lloc es triangulen totes les possibles combinacions de les mesures 2D de les imatges per

obtenir les posicions 3D candidates de cada *end-effector*. A continuació per cada *end-effector* es selecciona el candidat que està més aprop de la posició predita per el filtre d'estimació.

2.4 Reconstrucció de la postura

Com s'ha explicat al principi d'aquest capítol, en aquest treball es modela l'usuari utilitzant una cadena cinemàtica. Utilitzant cinemàtica inversa (IK), s'estima la configuració de la cadena cinemàtica a partir dels *end-effectors*, obtinguts en la fase de visió. D'aquesta manera s'obté la posició 3D de les articulacions de l'usuari per cada instant de temps. En aquesta secció s'explica la cinemàtica inversa i el model utilitzat per reconstruir la postura de l'usuari.

2.4.1 Cinemàtica inversa

Amb l'objectiu de capturar el moviment humà, el cos es modela com una cadena cinemàtica, que consisteix en un conjunt d'objectes rígids anomenats segments, connectats mitjançant articulacions rotacionals (veure Figura 2.10), on la seva configuració es descriu mitjançant un angle escalar. Encara que en aquest treball només s'utilitzin articulacions rotacionals, els algorismes i tota la teoria es pot aplicar a qualsevol tipus d'articulació [20].

La configuració completa d'una cadena cinemàtica ve donada pels escalars $\theta_1, \ldots, \theta_n$ que descriuen la configuració de les articulacions. Si s'assumeix que hi ha n articulacions, cada valor θ_j és el que s'anomena *angle de l'articulació* j. També es poden

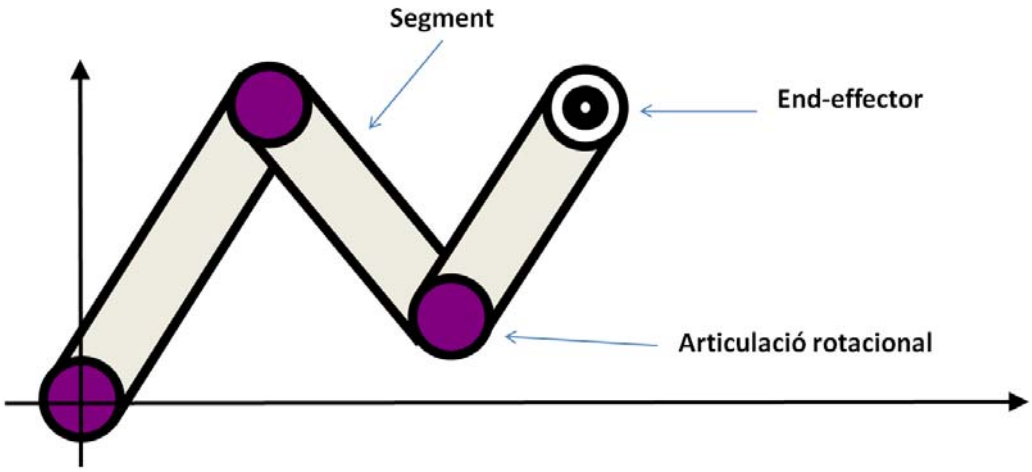

Figura 2.10: Exemple d'una cadena cinemàtica en el pla, que utilitza articulacions rotacionals.

especificar els angles de les articulacions com un vector columna $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$. Endemés de les articulacions, també hi haurà un cert nombre de punts de la cadena cinemàtica que correspondran als *end-effectors*. Si existeixen k *end-effectors*, les seves posicions seran descrites per $x_1, \ldots, x_k$. El vector $\mathbf{x}$ és la transposada del vector $(x_1, \ldots, x_k)^T$, que pot ser interpretat com un vector columna, ja sigui amb $m = 3k$ elements escalars o amb k elements de $\mathbb{R}^3$. La posició de cada *end-effector* x_i és una funció dels angles de les articulacions. El conjunt d'equacions no lineals que relacionen aquesta posició amb l'estat de les articulacions s'anomena el model geomètric directe. En aquest model no es considera cap tipus de moviment i es pot expressar com:

$$\boldsymbol{x} = f(\boldsymbol{\theta}) \tag{2.6}$$

Podem obtenir la posició del segment controlat a partir de la descripció de la configuració articular a partir de la matriu de transformació de cada articulació. En canvi, el problema invers de trobar la configuració de les articulacions a partir de la posició dels *end-effectors* es diu el model geomètric invers:

$$\boldsymbol{\theta} = f^{-1}(\mathbf{x}) \tag{2.7}$$

Invertir aquest sistema és possible sempre i quan la dimensió dels dos espais sigui la mateixa. Ara bé, si treballam amb cadenes articulades complexes on la dimensió dels dos espais no és la mateixa, aquesta inversió no és possible ja que per una configuració dels *end-effectors* hi ha més d'una configuració de les articulacions. En aquest cas s'ha de recórrer al model cinemàtic.

El model cinemàtic directe es basa en l'avaluació de variacions instantànies de les posicions dels *end-effectors* de la cadena per a cada articulació individual del sistema articular. D'aquesta manera es fa una linealització del model geomètric com es mostra a la Figura 2.11 (s'utilitza una analogia unidimensional per motius de claredat). S'ha de considerar que la linealització únicament és vàlida en un entorn de l'estat actual del sistema i que cada variació desitjada ha de verificar la hipòtesi de petits moviments.

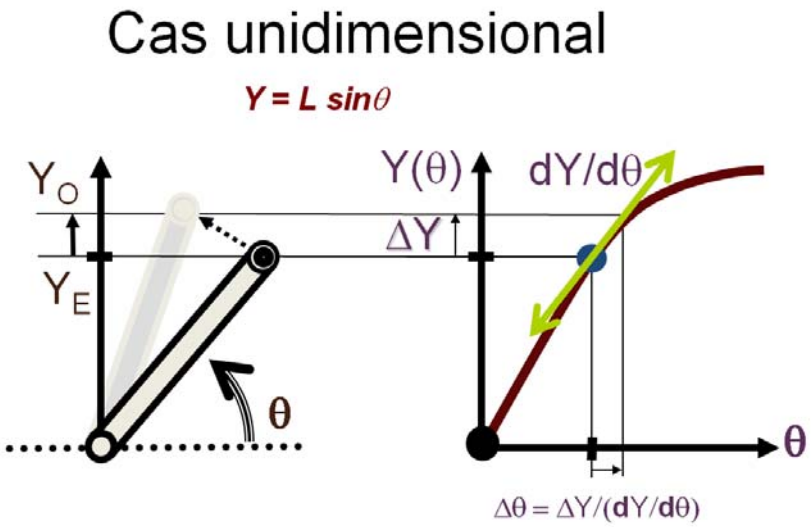

Figura 2.11: Linealització del model geomètric. En aquesta figura es considera el moviment de l'*end-effector* representat per la funció $Y(\theta)$ on $dY/d\theta$ és la derivada de Y respecta de θ.

Aquest esquema considera petits desplaçaments entorn a la configuració actual:

$$\mathbf{\Delta x} = J\mathbf{\Delta \theta} \tag{2.8}$$

On J és la matriu jacobiana $m \times n$ del sistema amb les variacions de primer ordre:

$$J_{i,j} = \frac{\partial x_i}{\partial \theta_j}; i = 1, \ldots, m; j = 1, \ldots, n; \tag{2.9}$$

En cas de que J sigui quadrada i no singular, llavors $\boldsymbol{\theta}$ es pot calcular a partir de:

$$\mathbf{\Delta \theta} = J^{-1}\mathbf{\Delta x} \tag{2.10}$$

S'obté la variació angular que ens du a la variació desitjada del segment final, és

el que es coneix com a cinemàtica inversa. Si el sistema és redundant, llavors n és major que m i el sistema no es pot invertir. Encara que no existeixi la matriu inversa i matemàticament no hi hagi solució per $\boldsymbol{\theta}$ podem utilitzar la inversa generalitzada per obtenir una resposta útil en aquests casos. La més utilitzada és la matriu pseudo-inversa J^+ que es pot calcular utilitzant el teorema de la descomposició en valors singular (SVD).

La SVD d'una matriu Jacobiana $m \times n$ de rang r [55] és:

$$J = \sum_{i=1}^{r} \sigma_i u_i v_i^T \tag{2.11}$$

on σ_i són els valors singulars (estrictament positius), $\{u_i\}$ i $\{v_i\}$ són les bases que s'estenen, respectivament, pel rang de l'espai de J i per l'espai complementari de $N(J)$.

L'expressió de la pseudo-inversa J^+ mostra la forta influència de qualsevol petit valor singular, la qual cosa explica la inestabilitat de la solució al voltant d'una singularitat:

$$J^+ = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T \tag{2.12}$$

La solució, presentada en [39, 46], consisteix en introduir un factor λ d'amortiment transformant el mal comportament del terme invers de l'equació anterior en un terme d'amortiment que convergeix suaument a zero quan un valor singular se converteix en petit:

$$J^{+\lambda} = \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2 + \lambda^2} v_i u_i^T \tag{2.13}$$

Quan λ és zero, es redueix a l'Equació 2.12. Per construcció, l'amortiment dels

mínims quadrats de la inversa $J^{+\lambda}$ defineix la qualitat de les restriccions de satisfacció per un límit alt a la solució.

Usant la pseudo-inversa $J^{+\lambda}$ la norma de la solució per $J^{+\lambda}$ és mínima, això és la variació de la postura més petita realitzant la variació desitjada:

$$\boldsymbol{\Delta\theta} = J^{+\lambda}\boldsymbol{\Delta x} \tag{2.14}$$

Mentre el $rank(\boldsymbol{J}) = m < n$ hi ha infinit nombre de solucions. Per al posicionament i l'animació de figures articulades en gràfics per ordinador, l'estratègia del pes [78] és freqüentment utilitzada per intentar limitar aquest nombre de solucions. En el camp de la robòtica, s'intenta resoldre la redundància afegint una tasca secundària a l'Equació 2.14, amb l'objectiu de minimitzar un criteri $h(\boldsymbol{\theta})$ sempre que sigui possible i no destorbi l'assoliment de la tasca principal. En aquesta formulació, la solució a la redundància s'aconsegueix movent les articulacions de tal manera que els *end-effectors* es desplacen cap a la posició desitjada i al mateix temps el criteri h és mantengui mínim. Aquesta idea va ser utilitzada per primera vegada per Liégeois [38] qui va afegir una tasca secundària projectant el gradient negatiu de $h(\boldsymbol{\theta})$ dins la projecció de l'espai nul $P_{N(J)}$, representada per l'equació:

$$\Delta\boldsymbol{\theta} = J^{+\lambda}\Delta\mathbf{x} - \alpha P_{N(J)}\nabla h(\boldsymbol{\theta}), \tag{2.15}$$

on α és un factor de guany positiu que depèn de la configuració. La definició de la tasca secundaria a través del criteri $\nabla h(\boldsymbol{\theta})$ depèn de l'aplicació. Per definició, l'espai nul del jacobià $N(J)$ s'assigna per J en el vector nul de l'espai restringit de

variacions. Dit més planerament, la variació del vector a través de $N(J)$ no té efectes sobre les restriccions. L'equació 2.16 correspon a la projecció de l'espai nul

$$P_{N(J)} = I_n - J^+ J, \tag{2.16}$$

on I_n és la matriu identitat $n \times n$.

A partir d'aquesta idea, Baerlocher et al. [2], generalitzen la cinemàtica inversa a p tasques o prioritats, cinemàtica inversa prioritzada (PIK). L'algorisme es basa en la linealització d'un conjunt d'equacions, expressant restriccions Cartesianes $\mathbf{x}$ com funcions dels graus de llibertat $\boldsymbol{\theta}$ de les articulacions. Es descriu la matriu Jacobiana $\mathbf{J}$ i s'usa la seva pseudo-inversa, descrita com $\mathbf{J}^+$, per construir la projecció dels operadors dins el nucli de $\mathbf{J}$, indicat com $N(\mathbf{J})$. L'algorisme del PIK es basa en el càlcul eficaç dels operadors de projecció que permeten dividir el conjunt de restriccions en múltiples subconjunts de restriccions associades amb un estricte nivell de prioritat [2]. La solució garanteix que una restricció associada amb una prioritat alta es durà a terme tan com sigui possible, mentre que una restricció amb prioritat baixa només serà optimitzada en el reduït espai de solucions que no pertorbi les restriccions amb major prioritat. Per tant, és molt important identificar quines restriccions tenen el major impacte sobre la qualitat de la convergència i l'aparença visual de la reconstrucció de la postura. Per exemple, el PIK és particularment adequat per l'avaluació *off-line* de l'espai assolible d'un treballador virtual; en aquest context la restricció de l'equilibri és la prioritat més alta mentre que la mirada i les posicions assolibles per l'usuari tenen un nivell de prioritat més baix [7].

La Figura 2.12 presenta un resum de l'esquema de control del PIK. La con-

vergència del bucle exterior és necessari ja que la linealització només és vàlida al voltant de l'estat actual; un domini de validesa tan petit requereix limitar la norma de qualsevol variació de la restricció desitjada $\Delta\mathbf{x}$ cap als seus respectius objectius com un valor màxim i iterar el càlcul de la solució prioritzada fins que la restricció s'hagi complida o que la suma dels errors assoleixi un valor constant. La Figura 2.12 també posa de relleu que el bucle de *clamping* tracta de la desigualtat associada dels límits mecànics de les articulacions. Bàsicament, es comprova si la solució prioritzada calculada $\Delta\boldsymbol{\theta}$ comporta la violació d'un o més límits. Si és el cas, la restricció d'igualtat s'insereix per fixar l'articulació marcada al seu límit i una nova solució prioritzada es cercada en l'espai reduït de l'articulació [2, 10].

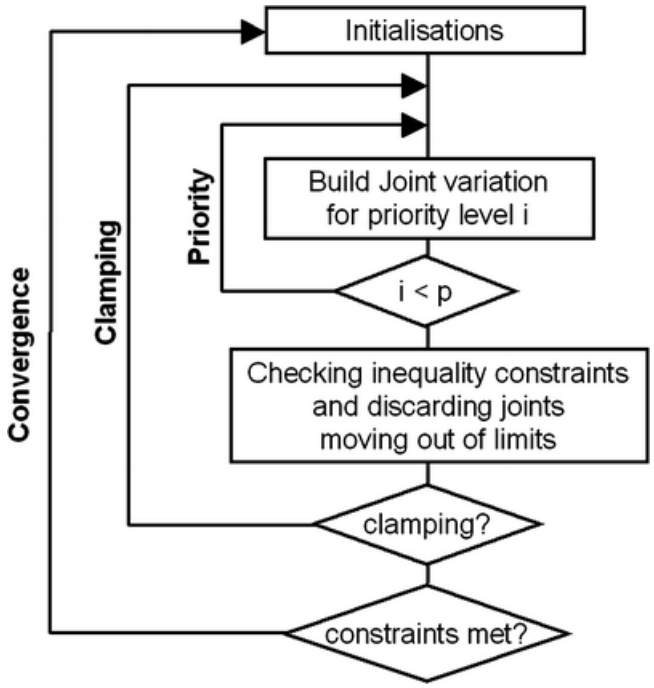

Figura 2.12: Esquema PIK. El bucle exterior itera la construcció de la solució de primer ordre amb prioritats i el bucle interior els límits de les articulacions.

2.4.2 Model i prioritats

Per recuperar la postura de l'usuari és important definir un model de l'usuari, que sigui senzill i suficient per analitzar els seus moviments. En aquest treball s'utilitza, com a model d'usuari, una cadena cinemàtica amb 15 graus de llibertat (dof en anglès) que és suficient per analitzar els seus moviments, com es va demostrar a [9]. Concretament, el model del cos de l'usuari està definit per un *Peu Virtual* (2 dofs), que connecta el cos al terra amb els eixos frontal i lateral de rotació, una *Esquena* (2 dofs), que correspon al principi de la columna amb els eixos frontal i lateral de rotació, el *Tòrax* (3 dofs), que té tots els eixos de rotació, les *Espatlles* (2×3 dofs) i els colzes (2×1 dof), veure Figura 2.13. El model s'inicialitza utilitzant una localització manual de les articulacions de les espatlles, els colzes i les mans amb l'objectiu de calcular la longitud de les extremitats, que seran constants per tota la captura. Per obtenir la posició inicial de la resta d'articulacions, es fa calculant la proporció relativa de la part inferior del segment del cos i el segment de l'esquena que són considerats constants. Així i tot durant la realització d'aquest treball s'ha estudiat la possibilitat de realitzar una inicialització del model automàtica, la qual s'exposa a l'Apèndix A.

Al treballar amb la recuperació de la postura d'una persona de peu, la realitat de la postura recuperada es regeix per la correctesa del seu equilibri. Per això el model proposat és capaç de modelar una distribució de massa simple de tot el cos i capaç d'oferir un control del centre de massa de tot el cos. I és per això que la prioritat més alta és la restricció del centre de massa, aquesta restricció assegura que el centre de massa es projectarà per sobre el node arrel (el Peu Virtual en la Figura 2.13) per

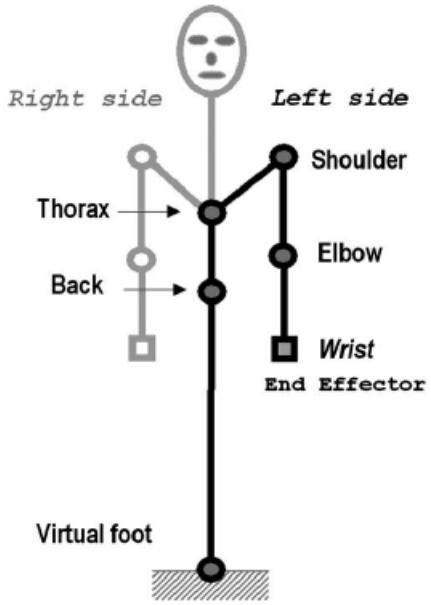

Figura 2.13: Model del cos de l'usuari.

garantir l'equilibri. A continuació, la següent restricció més important és la posició de la mà mitjançant el sistema de visió. Per cada mà, totes les articulacions des del colze fins al peu virtual contribueixen a l'assoliment d'aquesta restricció. Immediatament sota la restricció de la mà s'activen dues restriccions de baix nivell respectivament sobre les espatlles (atreta per la posició inicial en l'espai que es va obtenir en la fase de calibratge) i sobre els colzes (atrets cap a la seva posició més baixa possible per produir una postura més natural).

Per tant, es treballa amb 4 nivells de prioritats (veure Taula 2.2), per assegurar no només la posició dels *end-effectors* sinó també la propietat general que la cadena cinemàtica ha de garantir l'equilibri. Aquest esquema té dues conseqüències: la primera, és que permet evitar mínims locals que en un entorn sense prioritats d'una

manera o d'una altra s'haurien produït, i la segona, és que al ser equilibrades les postures intermèdies sempre seran més ben acceptades per l'espectador encara que la resta de restriccions no s'hagin complit. Això és important en un entorn de temps real, ja que pot passar que per cada conjunt d'imatges només es tengui temps per realitzar uns pocs passos de convergència del IK.

Restricció	Prioritat	dof's
Equilibri	1	2
Posició dels *end-effectors*	2	$2 \times 3 = 6$
Posició espatlla	3	$2 \times 3 = 6$
Posició colze	4	$2 \times 3 = 6$

Taula 2.2: Jerarquia de les restriccions prioritzades.

2.5 Avaluació del sistema de captura del moviment

En aquesta secció es presenta l'avaluació del sistema de captura del moviment de l'usuari. En primer lloc, es presenta l'entorn de captura que permet a l'usuari realitzar gestos naturals. A continuació, es mostren els resultats de l'avaluació realitzada per demostrar que el sistema treballa en temps real. Després de l'avaluació del temps real, es mostren els resultats de l'avaluació de la localització de les mans. Finalment, es mostra l'avaluació del sistema complet de captura.

2.5.1 Entorn de captura per gestos naturals

Un dels objectius principals del treball que es presenta en aquesta memòria, és que l'usuari ha de poder interaccionar realitzant gestos naturals. Aquest objectiu justifica el fet que els moviments es capturin mitjançant càmeres, ja que és un sistema no invasiu i per tant no limita l'usuari a l'hora de realitzar moviments. A l'hora de realitzar la captura dels moviments de l'usuari, s'ha definit un entorn de captura controlat, on aquest es situa en un espai entre la pantalla de projecció i les càmeres. Aquesta configuració, veure Figura 2.14, permet que l'usuari vegi la pantalla de projecció mentre realitza els moviments.

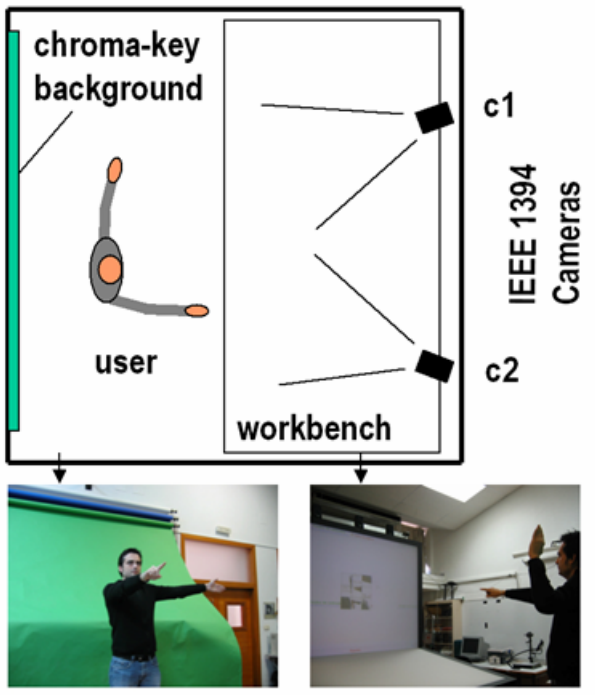

Figura 2.14: Disposició del sistema de visió.

Endemés de definir l'espai de captura, també s'han definit una sèrie de restriccions de cara a facilitar la captura del moviment. Dins l'espai de captura només hi pot estar una persona, ja que el sistema només pot capturar els moviments d'un individu. Per altra banda, el color de la roba de l'usuari no pot ser similar al color de la seva pell, i les parts del cos amb color de pell que no siguin les mans i la cara, no han de ser visibles (l'usuari no pot vestir roba màniga curta). Finalment, el fons ha d'estar cobert amb material chroma-key, encara que el sistema pot funcionar sense, però si s'utilitza s'assegura la resposta en temps real.

2.5.2 Rendiment

El sistema ha estat implementat en Visual C++ usant les llibreries OpenCV [12] i ha estat provat en un context d'interacció en temps real sobre un AMD Athlon 2800 + 2.083 GHz amb Windows XP. Les imatges han estat capturades usant dues càmeres DFW-500 de Sony. Les càmeres proporcionen imatges de 320×240 amb un rati de captura de 30 imatges per segon. En les proves de laboratori s'ha descobert que el sistema opera a 48Hz (24 fps per cada càmera) si no s'itera el PIK. Si s'usen 5 interaccions el sistema treballa a 22 fps i per un màxim de 20 iteracions el sistema opera a 19 fps. Aquests resultats garanteixen la resposta en temps real del sistema.

2.5.3 Localització de les mans

L'algorisme de visió per ordinador s'ha validat mesurant l'exactitud dels resultats, o sigui la posició 3D dels *end-effectors*. La posició 3D es localitza amb un dispositiu de posicionament per ultrasons, l'IS-900 MiniTrax Wireless Hand de InterSense

Figura 2.15: Configuració per avaluar l'algorisme de visió per ordinador.

Company. En aquest experiment, l'usuari sosté el dispositiu amb una mà, veure Figura 2.15. Llavors, s'obté un seguiment de les posicions estimades pel sistema presentat i de les posicions informades pel dispositiu IS-900 al mateix instant de temps. Amb l'objectiu d'avaluar quantitativament, es calcula l'error quadràtic (mitjà) a partir dels dos conjunts de punts en el mateix sistema de referència Formalment, l'error entre una posició 3D d'una articulació estimada X^e i la posició verdadera captura pel dispositiu X^{GT} es calcula com:

$$E(\boldsymbol{X}^e, \boldsymbol{X}^{GT}) = \frac{1}{i} \sum ||\boldsymbol{X}^e - \boldsymbol{X}^{GT}||_2 \tag{2.17}$$

on i són el nombre de punts i $|| \cdot ||_2$ és la norma euclidiana.

Amb l'objectiu de fer experiments exhaustius es duen a terme un conjunt de diferents experiments:

- Comparació entre posicions claus estàtiques

- Comparació de moviments predefinits ("moviment del braç")
- Comparació de seqüències curtes de moviments aleatoris
- Comparació de seqüències llargues de moviments aleatoris

Experiment	E (in mm)	Nombre de frames
Estàtic	4.8	376
Moviment del braç	12.4	116
Moviments aleatoris (curt)	40.3	849
Moviments aleatori (llarg)	54.3	2465

Taula 2.3: Resultat d'avaluació dels seguiment 3D dels *end-effectors*.

La Taula 2.3 mostra l'error mitjà obtingut en les diferents proves amb diferents usuaris per quatre experiments. Primer, l'experiment amb una posició estàtica és útil per mesurar l'error de jitter dels dos dispositius, que pot ser quantificat en 4 mil·límetres (de fet, aquest valor és la precisió mínima del sensor d'ultrasons d'InterSense). En els experiments, es pot veure que l'error mig augmenta i s'estabilitza amb un màxim de 55 mm. L'únic requeriment d'aquest experiment és que les dues mans han de ser completament visibles en les dues imatges. Per exemple quan l'usuari té els seus braços completament estesos. Per complir aquest requeriment, en els experiment s'utilitza una lent de càmera de 6mm, de manera que l'usuari ha d'estar col·locat en un rang de distàncies d'entre 2.5m i 3.5 m des del parell estèreo de càmeres. De fet, entre 3 i 3.5 metres, l'usuari pot fer qualsevol moviment assegurant que les seves mans seran visibles per les dues càmeres. Dins aquest rang de distancies, l'error no varia significativament. Aleshores, d'acord amb els resultats obtinguts

en aquests experiments, l'error principalment és degut a la forma de la mà. La mà és capturada des de les càmeres en diferent orientacions, d'aquesta manera el punt de referència per localitzar la mà (el centre de gravetat) varia amb la seva forma. Aquesta és la desviació principal dels mesuraments del dispositiu d'ultrasons. En la Figura 2.16, són mostrats els seguiments en l'espai 3D pels dos posicionament del sistema en dos experiments diferents. Es pot veure en aquestes figures que el seguiment és igual a algunes desviacions degut a les diferents formes de la mà capturades.

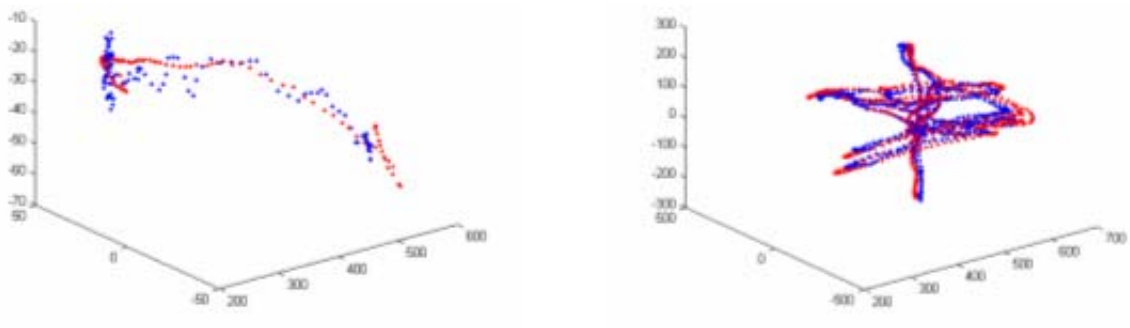

Figura 2.16: Esquerra: trajectòries 3D d'un moviment predefinit. Dreta: Trajectòries 3D d'un moviment aleatori. Sensor d'ultra sons en vermell, sistema de captura en blau.

2.5.4 El sistema complet

Amb l'objectiu d'avaluar el sistema complet incloent el PIK, s'han comparat els resultats de l'aplicació envers les posicions reals utilitzant dues seqüències anotades [51]. S'ha comparat la posició dels colzes entre els punts anotats i els detectats. Per a la comparació, s'han triat les posicions dels colzes perquè són les articulacions de la part superior del cos humà que en aquestes dues escenes els seus valors són

estimats mitjançant la combinació del seguiment dels *end-effectors* guiat per la visó i l'estimació de l'articulació del PIK. La primera seqüència té 450 imatges que corresponen a 15 segons de temps real. En aquesta seqüència, els moviments humans són suaus i no hi ha oclusions difícils entre els *end-effectors* que puguin distreure el procés de captura. En aquest test, l'error mitjà de l'estimació d'ambdós colzes envers les posicions reals és similar i pot ser quantificada al voltat de **50 mm**. La segona seqüència és composada per 600 imatges, que corresponen a 20 segons de temps real. En aquesta seqüència l'usuari mou els seus braços lliurement sense cap restricció. Els moviments són ràpids i existeixen oclusions importants dels *end-effectors*, per exemple quan l'usuari creua els seus braços, veure Figura 2.17. En aquest cas, l'error produït per ambdós és també similar i pot ser quantificat al voltant de **120 mm**. L'error pot ser alt si l'usuari aixeca el colze perquè el PIK atreu el colze per avall perquè suposam que aquest és més natural i no es disposa de cap altra informació de control del colze.

Figura 2.17: Segona seqüència de prova. En aquesta seqüència l'usuari mou els braços sens cap restricció, lliurement. Es produeixen oclusions dels *end-effectors*.

Finalment, també s'ha provat l'aplicació realitzant varis moviments de braç predefinits i comparant els resultats amb les posicions finals desitjades entre moviments, amb l'objectiu de que la jerarquia funcioni correctament, la posició inicial del braç de l'usuari ha de ser completament estesa al llarg del cos per així poder determinar l'extensió màxima del braç. En primer lloc, l'usuari ha de flexionar un colze fins a la màxima flexió (això no és fàcil per la cinemàtica inversa perquè la postura inicial és singular); en segon lloc, el centre de massa influent pot ser provat usant només una articulació espatlla per moure el braç lateralment: quan el braç és horitzontal tracta d'assolir el punt més lateral. Això obligarà a l'usuari a contrarestar la postura de la part superior del cos amb la part inferior. Com que la prova del colze i la tasca del centre de massa funciona correctament en aquests casos es pot veure a la Figura 2.18.

Figura 2.18: Postures estimades de diferents moviments predefinits del braç.

2.6 Resum

En aquest capítol s'ha presentat un mètode basat en visió que permet obtenir els moviments de l'usuari en un espai 3D. A partir d'un parell d'imatges estèreo, algorismes de visió per ordinador localitzen la posició 3D de les mans de l'usuari. A partir d'aquesta localització de les mans, algorismes de cinemàtica inversa amb prioritats estimen la postura de l'usuari mitjançant una cadena cinemàtica de 15 graus de llibertat.

El principal avantatge d'aquest sistema, és que evita específicament mètodes intrusius tal com són els marcadors, endemés de permetre a l'usuari realitzar un àmplia gamma de moviments.

Els resultats de l'avaluació, mostren que la qualitat dels resultats és suficient per l'objectiu proposat, que consisteix en realitzar la reconstrucció de la postura de l'usuari per a la interacció 3D en temps real.

Capítol 3

Reconeixement de gestos per a la interacció natural

Generalitzar sempre és equivocar-se.

Hermann von Keyserling.

En aquest capítol es descriu com es reconeixen els gestos de l'usuari a partir de les postures capturades. Per fer-ho, s'utilitzen les posicions 3D de les articulacions de la cadena cinemàtica que modela la postura de l'usuari en cada instant de temps. A partir d'aquestes posicions es resolen els principals problemes del reconeixement de gestos que són les variacions temporals, espacials i d'estil. A més, un punt important del reconeixement de gestos que es presenta és que cada usuari realitza la interacció utilitzant els gestos que considera més adients.

3.1 Enfocament

La comunicació no-verbal és habitualment entesa com el procés en què les persones enviam i rebem missatges no orals [35]. De fet la majoria de comunicació que rebem és no-verbal, com són els gestos, la postura, la forma de vestir, la forma de mirar,... Concretament, un gest és una forma de comunicació no verbal feta amb una part del cos, que a vegades s'usa conjuntament amb la comunicació verbal.

El llenguatge dels gestos, o sigui la comunicació mitjançant gestos, és suficientment rica per expressar gran varietat de missatges, encara que normalment el significat d'un gest depèn de l'experiència i l'origen del transmissor i/o del receptor, de la mateixa forma que existeixen diferents idiomes. Per exemple, és ben conegut que els italians i les italianes tenen un bon conjunt de gestos d'origen cultural, veure Figura 3.1. Alguns d'aquests gestos tenen significat per nosaltres, ja que és una cultura propera, altres no. I segurament, no tots els gestos tenen el mateix significat a tota a Itàlia.

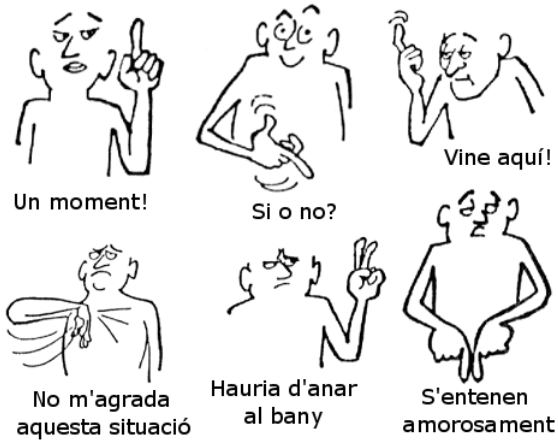

Figura 3.1: Alguns exemples de gestos culturals italians.

Per altra part, alguns gestos han estat creats per les persones i el seu significat està regulat. En aquest grup s'hi pot incloure el llenguatge de signes que utilitzen les persones sordmudes, del qual n'existeix un per cada idioma. Com també s'hi pot incloure els gestos que utilitzen els agents de tràfic, que estan regulats en el codi de circulació. La Figura 3.2 mostra un exemple de gest del codi de circulació i un exemple de l'alfabet en llenguatge de símbols.

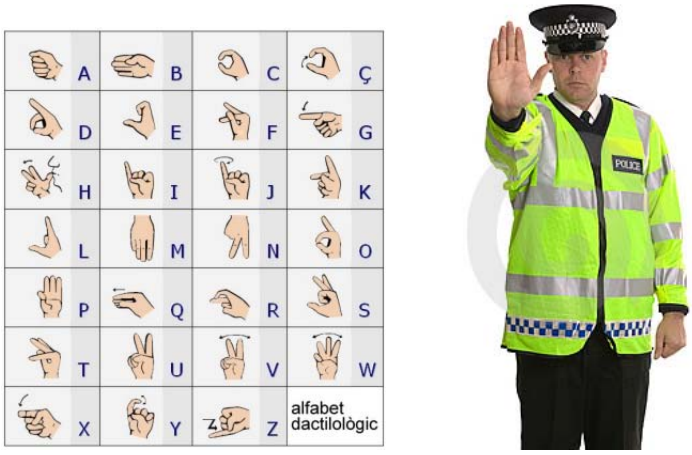

Figura 3.2: Esquerra, alfabet en llenguatge de símbols. Dreta, gest del codi de circulació que significa aturar.

Pel fet que el significat de la majoria de gestos existents depenen de l'origen de la persona que els utilitza, en aquest treball es proposa que l'usuari pugui utilitzar els gestos que consideri més naturals a l'hora d'interaccionar. Així l'usuari podrà centrar-se en la interacció amb el sistema, i no s'haurà de preocupar d'aprendre un conjunt de gestos, que tal vegada per ell no siguin naturals, ni haurà de dedicar part de la concentració durant la sessió a utilitzar els gestos predefinits.

Per això, a partir de les posicions 3D de les articulacions de l'usuari, que ens

proporciona el sistema de captura presentat en el Capítol 2, es solucionen els principals problemes del reconeixement de gestos, que s'han observat durant la realització d'aquest treball, són les variacions temporals, les variacions espacials i les variacions d'estil.

Les variacions espacials són degudes a les restriccions físiques del cos humà com poden ser les diferents talles dels usuaris. Per resoldre-les, el reconeixement de gestos ha de ser invariant a les variacions espacials. Dit d'una altra manera que per molt diferents que siguin dos usuaris si fan el mateix gest, el sistema l'ha de reconèixer com el mateix.

Per altra banda, les variacions temporals són degudes a les diferents velocitats en què els usuaris duen a terme els gestos. Aquestes són gestionades utilitzant una representació temporal del gest. Independentment de la velocitat en què l'usuari realitzi els gestos aquest s'han de reconèixer.

Finalment, les variacions d'estil són degudes a la forma personal en què cada usuari realitza el seus gestos. Aquest, és el repte més difícil i més important perquè el sistema presentat treballi amb gestos naturals. Per resoldre aquestes variacions es parametritzen els gestos de cada usuari mitjançant una fase d'aprenentatge en l'inici de cada sessió.

3.2 Treballs previs

Com s'ha comentat a la introducció, la idea d'utilitzar gestos per interaccionar amb l'ordinador no és nova, ja el 1980 Bolt va presentar una interfície multimodal [6] que

va anomenar *Put That There*, on combinava el reconeixement de veu i l'assenyalament per moure objectes d'una escena.

La tecnologia de la visió per computador, aplicada a les interfícies persona-ordinador, està tenint actualment un notable èxit [44, 42, 43, 75]. Des del punt de vista de la interacció persona-ordinador, es interessant reconstruir els moviments de l'usuari amb la finalitat de reconèixer els gestos, que el sistema pot interpretar com esdeveniments. En aquest sentit, els enfocaments utilitzats per al reconeixement i l'anàlisi del moviment humà en general es poden classificar en tres grans categories: les basades en el moviment, les basades en l'aparença i les basades en el model. Les propostes basades en el moviment intenten reconèixer el gest directament des del moviment sense utilitzar cap informació estructural del cos de l'usuari [53, 4, 23]. Les propostes basades en l'aparença usen informació bidimensional tal com imatges en escala de grisos, contorns o siluetes del cos [64, 24]. En canvi, les propostes basades en el model es centren en recuperar la configuració tridimensional de les parts del cos articulat [57, 72, 59, 36]. No obstant, les propostes basades en model són habitualment les més difícils d'aplicar en aplicacions del món real. Bàsicament és degut a la dificultat de capturar i seguir les articulacions del cos que formen part dels gestos.

Una solució parcial pot ser simplificar la captura a poques parts del cos i usar les seves trajectòries temporals amb l'objectiu de reconèixer els gestos d'interès [76]. Per exemple, Rao et al. [56] consideren el problema d'aprendre i reconèixer accions realitzades per una mà humana. Per fer-ho, utilitzen la invariància afí i apliquen el seu mètode en imatges d'una seqüència real utilitzant el color de pell per trobar

les mans. A continuació, caracteritzen un gest a partir dels moments dinàmics, que es defineixen com la màxima curvatura espacio-temporal de la trajectòria de la mà, que es conserva quan es transforma del 3D al 2D. El seu sistema no requereix un model, de fet es va construint una base de dades de models memoritzant els gestos que els usuaris realitzen. Una altra proposta de reconeixement de gestos basat en les mans usa la postura de la mà com a gest per navegar en mons virtuals [48]. No obstant, utilitzar una sola localització 3D d'una o dues mans no és suficient per al reconeixement de gestos complexes per controlar aplicacions interactives.

En el treball de Polana et al. [54] es mostra que el moviment humà pot ser reconegut utilitzant representacions no paramètriques de baix nivell, de fet demostra que un moviment repetitiu és una senyal tan forta que l'actor en moviment pot ser segmentat, normalitzat espacialment i temporal, i reconegut comparant-lo amb un patró espacio-temporal de característiques del moviment. En canvi, en el treball de Efros et al. [23] es proposa un descriptor del moviment basat en el mesurament del flux òptic en un volum espacio-temporal per cada figura humana estabilitzada, i una mesura de similitud associada per ser usada en un entorn del veí més proper. Per classificar l'acció duta a terme per l'usuari, utilitza el veí més proper amb una base de dades de seqüències de vídeo ja classificades. Bobick i Davis [4] deriven la representació del patró temporal amb imatges amb el fons eliminat. Presenten resultats sobre diverses accions duites a terme per diferents usuaris. Freeman et al. [26] usa els moments de la imatge i histogrames d'orientació del gradient de la imatge, per interactuar amb el control d'un videojoc. Són solucions bones si disposam d'imatges de baixa resolució i en casos que és molt complicat obtenir el model de

l'usuari. Totes aquestes propostes treballen a partir de l'aparença de la imatge, les quals obtenen bons resultats per al control d'accions repetitives fetes amb tot el cos (caminar, córrer,...), però no per al control de gestos específics realitzats amb algunes parts del cos. Tal vegada per dur a terme el reconeixement de gestos seria una bona idea implementar una solució intermèdia, on a partir de l'aparença com entrades parcials és reconstruís el model 3D de l'usuari.

Per altra banda, en la majoria dels enfocaments per reconèixer gestos utilitzen els Hidden Markov Models (HMM), per tractar estadísticament les propietats temporals dels gestos, en gran part degut al fet d'usar directament els valors de les imatges [75]. Aquests enfocaments no s'apliquen en temps real, perquè els HMM requereixen un important fase d'aprenentatge per poder ajustar tots els paràmetres del model.

3.3 Representació de la postura

Com s'apuntava a la Secció 3.1, les variacions espacials són un dels principals problemes a resoldre a l'hora de reconèixer els gestos de l'usuari, un exemple són les diferents talles dels usuaris. El sistema de captura retorna la modelització de la postura de l'usuari, mitjançant una cadena cinemàtica, per cada instant de temps. Per tant, el que realment s'ha de trobar és una representació d'aquesta postura que sigui espacialment invariant.

Una de les primeres diferencies espacials, que presenta el sistema de captura, té a veure amb el procés de calibratge. En el procés de calibratge s'utilitza un patró pla, per calcular els paràmetres intrínsecs i extrínsecs del parell estèreo de càmeres [77].

Usant aquest esquema, el sistema de coordenades es col·loca respecte de l'objecte de calibratge, com es mostra a la Figura 3.3. Aleshores, el sistema de coordenades canvia cada vegada que es calibra el sistema, i les posicions de la cadena cinemàtica són referenciades des d'un origen del món desconegut, que dependrà d'on s'hagi situat el patró de calibratge.

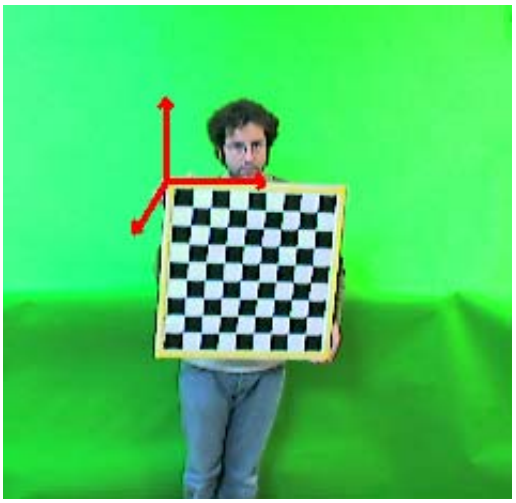

Figura 3.3: Sistema de referència en el procés de calibratge.

Per resoldre aquest problema, es proposa que el sistema de coordenades s'alineï automàticament amb la posició i orientació de l'usuari en el primer instant de temps, amb un canvi del sistema de coordenades, com es mostra a la Figura 3.4. En primer lloc, posicionant l'origen del sistema de referència sobre la posició del peu de l'usuari, mitjançant una translació. A continuació, alineant l'eix y amb el vector director que uneix el peu de l'usuari i la seva espatlla, mitjançant una rotació. Finalment alineant l'eix x amb el vector director que uneix l'espatlla esquerra i dreta de l'usuari, anul·lant el valor de la component y, mitjançant una rotació. El lector pot

observar, que aquestes són les operacions típiques per realitzar un canvi qualsevol d'un sistema de coordenades a un altre.

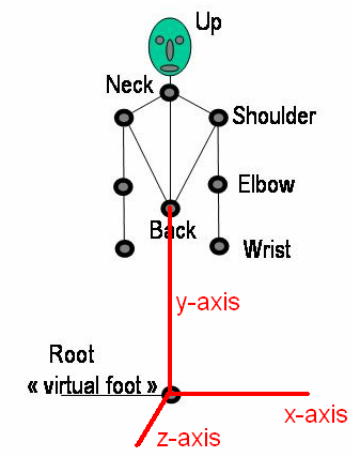

Figura 3.4: Alineament del sistema de referència amb l'usuari.

Amb aquest canvi del sistema de referència s'aconsegueix que tots els usuaris tenguin el mateix, i a més les posicions 3D de les articulacions són independents de l'entorn, perquè el sistema de referència és alineat amb el cos de l'usuari i no depèn del procés de calibratge.

L'altra variació espacial a tenir en compte, com s'ha explicat anteriorment, són les diferents talles dels usuaris. Una possibilitat per representar la postura independentment de la talla, és usant la informació del moviment de les articulacions mitjançant angles d'Euler [44]. No obstant, en aquest cas, la informació del moviment és inestable, és a dir, petits canvis d'aquests valors poden produir deteccions equivocades.

Como alternativa, es proposa una representació de cada segment del cos mitjançant un vector director, que representa l'orientació predominant del segment.

Formalment, el vector director $\vec{u}_l$, que representa l'orientació del segment l, definit per les articulacions $\mathbf{J}_1$ i $\mathbf{J}_2$, és calculat com es mostra a continuació

$$\vec{u}_l = \frac{\mathbf{J}_2 - \mathbf{J}_1}{\|\mathbf{J}_2 - \mathbf{J}_1\|}, \tag{3.1}$$

on $\mathbf{J}_i = (x_i, y_i, z_i)$ és la posició 3D de l'articulació i en el sistema de referència centrat a l'usuari. D'aquesta forma, depenent de l'alfabet de gestos desitjat, només serà necessari calcular els vectors directors dels segments implicats. Aquesta representació transforma les dades de forma que són invariants respecte la talla de l'usuari i es solucionen finalment les variacions espacials, ja que tots els segments de l'usuari i de qualsevol usuari tendran la unitat com a longitud.

Com s'ha explicat en el paràgraf anterior, a l'hora de reconèixer els gestos només serà necessari calcular els vectors directors implicats en el conjunt de gestos per reconèixer. En el cas de què l'usuari només utilitzés les mans, bastaria calcular i tenir en compte els vectors dels braços. Per tant, es proposa com idea per representar la postura del cos de l'usuari, utilitzar un vector de característiques compost pels vectors directors dels segments de l'usuari implicats en el conjunt de gestos. Formalment, la representació de l'orientació d'un segment, l, és

$$\mathbf{q}^l = (u_x^+, u_x^-, u_y^+, u_y^-, u_z^+, u_z^-), \tag{3.2}$$

on u_x^+ i u_x^- són respectivament la magnitud positiva i la negativa de la component x del vector director, u_x, i es compleix que $u_x = u_x^+ - u_x^-$ i $u_x^+, u_x^- \geq 0$. El mateix s'aplica per les components u_y i u_z. D'aquesta forma, les components d'orientació

del vector unitari d'un segment són *half-wave rectified* dins sis canals no negatius.

A partir de la representació de l'orientació de cada segment implicat, es construeix un histograma, per representar la postura com un tot i no com un conjunt de segments de l'usuari. En aquest treball de recerca es proposen dues formes per construir l'histograma, l'acumulada i l'enllaçada.

En la primera representació que es proposa, l'acumulada, l'histograma que representa la postura va acumulant les orientacions dels segments, com es mostra a continuació

$$\mathbf{q} = \frac{\sum_{l=1}^{n} \mathbf{q}^l}{\| \sum_{l=1}^{n} \mathbf{q}^l \|} \tag{3.3}$$

on n és el nombre de segments implicats en els gestos a reconèixer. Com ens indica la fórmula, l'histograma està normalitzat. Amb aquesta representació s'aconsegueix una representació de la postura amb un histograma de sis elements (bins), que és la dimensió de la representació de les orientacions de cada segment. Al acumular les orientacions i normalitzar-les pot passar que dues postures diferents tenguin la mateixa representació.

En la segona representació que es proposa, l'enllaçada, l'histograma que representa la postura va enllaçant les orientacions dels segments, com es mostra a continuació

$$\mathbf{q} = \{\mathbf{q}^l\}_{l=1..n}, \quad \sum_{l=1}^{n} \mathbf{q}^l = 1 \tag{3.4}$$

on n és el nombre de segments implicats en els gestos a reconèixer. Amb aquesta representació s'aconsegueix que cada postura tengui una representació única. En canvi la dimensió de l'histograma dependrà del nombre de segments utilitzats.

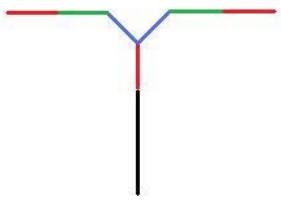

Figura 3.5: Postura ideal dels dos braços estesos.

A partir de les dues representacions proposades, es pot apreciar que la representació acumulada permet representar un conjunt de postures més reduïda. En el cas en què el mateix moviment es faci amb diferents segments, no es podran distingir, com és el cas de fer el mateix moviment amb els dos braços. A continuació es presenta un exemple ideal pas a pas de com representar la postura, de dos usuaris diferents que realitzen la mateixa postura, perquè l'exemple sigui més senzill només s'utilitzaran els dos avant braços per representar la postura. La postura que realitzarien els dos usuaris es pot veure a la Figura 3.5.

Imaginem que el sistema de captura, una vegada s'ha fet el canvi del sistema de referència, ha retornat les posicions 3D que es mostren a la Taula 3.1.

Articulació	Usuari 1	Usuari 2
colze esquerre	(43,149,0)	(52,162,0)
colze dret	(-43,149,0)	(-52,162,0)
canell esquerre	(60,149,0)	(70,162,0)
canell dret	(-60,149,0)	(-70,162,0)

Taula 3.1: Posicions 3D de les articulacions de l'usuari, per la postura d'exemple mostrada a la Figura 3.5.

A partir d'aquestes dades, es calculen els vectors directors de cada segment de cada usuari. Els resultats es mostren a la Taula 3.2. En aquesta Taula es pot apreciar que els vectors directors dels usuaris són els mateixos, ja que els dos usuaris realitzen la mateixa postura.

Segment	Usuari 1	Usuari 2
Esquerre	(1,0,0)	(1,0,0)
Dret	(-1,0,0)	(-1,0,0)

Taula 3.2: Vectors directors dels usuaris, per la postura d'exemple mostrada a la Figura 3.5.

La següent passa es generar els vectors de característiques de cada segment, que consisteix sis canals no negatius que representen les orientacions de cada vector unitari, veure Taula 3.3.

Segment	Usuari 1	Usuari 2
Esquerre	(1,0,0,0,0,0)	(1,0,0,0,0,0)
Dret	(0,1,0,0,0,0)	(0,1,0,0,0,0)

Taula 3.3: Vector de característiques de cada segment, per la postura d'exemple mostrada a la Figura 3.5.

Finalment, s'ha de representar la postura en una de les dues formes proposades, l'enllaçada o l'acumulada. Per a l'exemple que s'està realitzant, el resultat d'ambdues representacions es presenta a la Taula 3.4.

Representació	Usuari 1	Usuari 2
Enllaçada	(1,0,0,0,0,0,0,1,0,0,0,0)	(1,0,0,0,0,0,0,1,0,0,0,0)
Acumulada	(1,1,0,0,0,0)	(1,1,0,0,0,0)

Taula 3.4: Resultat d'aplicar la representació acumulada i enllaçada a la postura d'exemple mostrada a la Figura 3.5.

A continuació per veure que l'enllaça pot representar un conjunt major de postures anem a realitzar un altre exemple, el mateix usuari fa dues postures diferents, que són l'efecte mirall una respecta de l'altra, veure la Figura 3.6

Figura 3.6: Postures mirall.

Imaginem que el sistema de captura, una vegada s'ha fet el canvi del sistema de referència, ha retornat les posicions 3D que es mostren a la Taula 3.5.

Articulació	Usuari 1	Usuari 2
colze esquerre	(20,177,0)	(20,121,0)
colze dret	(-20,121,0)	(-20,177,0)
canell esquerre	(20,204,0)	(20,94,0)
canell dret	(-20,94,0)	(-20,204,0)

Taula 3.5: Posicions 3D de les articulacions de l'usuari, per la postura d'exemple mostrada a la Figura 3.6.

A partir d'aquestes dades, es calculen els vectors directors de cada segment de cada usuari. Els resultats es mostren a la Taula 3.6.

Segment	Usuari 1	Usuari 2
Esquerre	(0,1,0)	(0,-1,0)
Dret	(0,-1,0)	(0,1,0)

Taula 3.6: Vectors directors dels usuaris, per la postura d'exemple mostrada a la Figura 3.6.

La següent passa es generar els vectors de característiques de cada segment, que consisteix sis canals no negatius que representen les orientacions de cada vector unitari, veure Taula 3.7.

Segment	Usuari 1	Usuari 2
Esquerre	(0,0,1,0,0,0)	(0,0,0,1,0,0)
Dret	(0,0,0,1,0,0)	(0,0,1,0,0,0)

Taula 3.7: Vector de característiques de cada segment, per la postura d'exemple mostrada a la Figura 3.6.

Finalment, s'ha de representar la postura en una de les dues formes proposades,

l'enllaçada o l'acumulada. Per a l'exemple que s'està realitzant, el resultat d'ambdues representacions es presenta a la Taula 3.8. Es pot veure que la representació acumulada és la mateixa per les dues postures.

Representació	Usuari 1	Usuari 2
Enllaçada	(0,0,1,0,0,0,0,0,0,1,0,0)	(0,0,0,1,0,0,0,0,0,1,0,0,0)
Acumulada	(0,0,1,1,0,0)	(0,0,1,1,0,0)

Taula 3.8: Resultat d'aplicar la representació acumulada i enllaçada a la postura d'exemple mostrada a la Figura 3.6.

Un esquema resum de com es realitzen les dues representacions, es mostra a la Figura 3.7. A partir de les imatges capturades pel parell estèreo de càmeres, s'estima la postura de l'usuari amb la proposta presentada al Capítol 2. A continuació, de la configuració de la cadena cinemàtica retornada pel sistema de captura, s'extreuen les posicions de les articulacions que volem utilitzar per representar la postura. En el cas de l'esquema mostrat a la Figura 3.7, s'extreuen els colzes i els canells. A partir de la posició dels colzes i els canells, es calculen els vectors directors, que en aquest cas representen la direcció dels avantbraços. Seguidament, amb els vectors directors es calculen els vectors de característiques de cada segment. Finalment, a partir d'aquest vector de característiques es representa la postura, o bé acumulant aquests vectors o bé enllaçant-los.

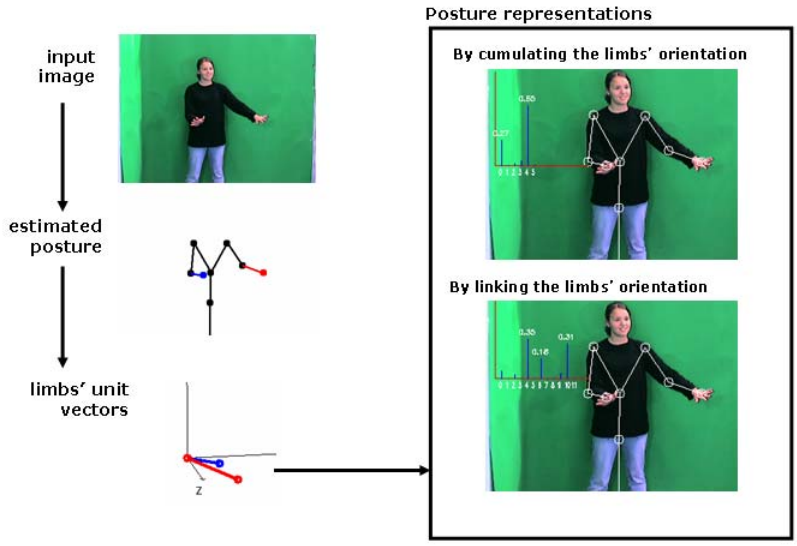

Figura 3.7: Construcció de la representació de la postura.

3.4 Representació del gest

Amb la representació de la postura que s'ha proposat a la secció anterior, s'aconsegueix que per molt diferents que siguin dos usuaris si fan el mateix gest, la representació sigui la mateixa. En aquesta secció es proposa una forma de solucionar les variacions temporals, dit d'una altra manera, independentment de les diferents velocitats en què els usuaris realitzin els gestos, aquests s'han de reconèixer.

La nostra proposta de representació del gest es basa en una representació temporal de la postura de l'usuari. La raó per usar informació de la postura és que les postures defineixen directament els gestos, fins i tot en molts de casos amb només una postura és possible reconèixer un gest, veure Figura 3.8.

Figura 3.8: Postura amb significat.

Encara que per assegurar-se que es tracte d'un gest, enlloc d'una posició de transició, és convenient representar el gest amb més d'una postura. En el cas de la imatge anterior (Figura 3.8), si la postura es mantén un cert temps es pot confirmar que el policia està realitzant un gest que significa aturar-se. En canvi si la postura només es mantén en una imatge, significaria que és una postura de transició. Per aquest motiu, en aquest treball es considera que un gest està compost per una seqüència de postures que realitza l'usuari, per tant un gest té una component temporal important que el diferencia d'una postura.

Per representar un gest, es proposa utilitzar un histograma que acumula la seqüència de postures que hi estan involucrades, per tant la definició matemàtica d'un gest és:

$$\hat{\mathbf{q}}_t = \frac{1}{T} \sum_{i=t-T}^{t} \mathbf{q}_i, \tag{3.5}$$

on t és l'instant actual i T és la periodicitat del gest, i pot ser interpretada com una finestra temporal de les postures acumulades. Aquest procés assumeix les variacions temporals dels gestos mitjançant una detecció de la periodicitat de cada interpretació del gest de cada usuari amb l'objectiu de fixar el valor T, això és, el seu grau temporal.

La Figura 3.9 mostra una seqüència d'imatges on un usuari realitza un gest, en aquest cas per a l'usuari el gest significa rotar, ja que quan el realitzava havia de rotar un objecte que apareixia en la pantalla.

Figura 3.9: Representació del gest acumulada.

A partir d'aquest gest la Figura 3.10 mostra l'histograma que el representa, a partir de la representació acumulada de la postura.

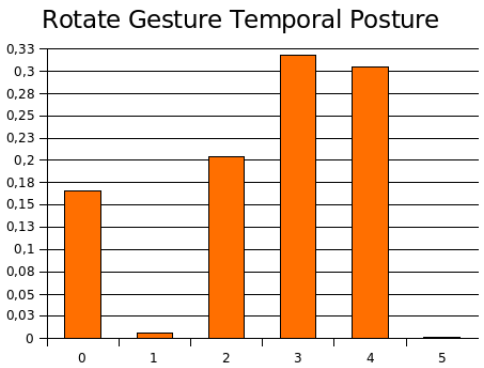

Figura 3.10: Representació del gest acumulada.

En canvi la Figura 3.11 mostra l'histograma que representa el gest, a partir de la representació enllaçada de la postura.

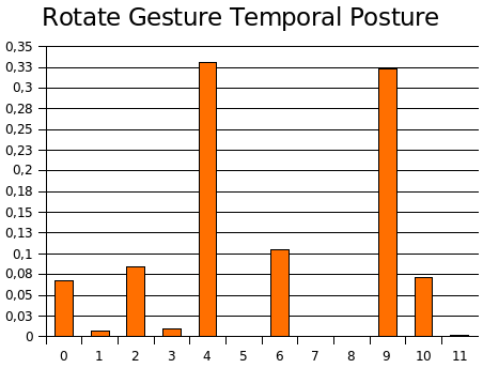

Figura 3.11: Representació del gest enllaçada.

3.5 Reconeixement del gest

Finalment falta resoldre les variacions d'estil. Com s'ha comentat a la Secció 3.1, són degudes a la forma personal en què cada usuari realitza el seus gestos. Un mateix usuari mai realitza el mateix gest exactament, sempre hi ha petites diferències encara que no sempre es percebin. De fet, les persones que utilitzen el llenguatge de signes, cada cert temps han de realitzar un reciclatge, per corregir els mals hàbits adquirits per l'acumulació de petites variacions d'estil que van realitzant.

Endemés, com s'ha mostrat en varis experiments amb nins [33], un gest és natural depenent de l'experiència de l'usuari. A l'hora d'interaccionar amb un joc, diferents usuaris utilitzaven diferents gestos per la mateixa funció d'interacció. Per aquesta raó, també és necessari que els usuaris puguin utilitzar els gestos que considerin més adients per cada tipus d'interacció.

Per aconseguir que el sistema resolgui les variacions d'estil i que l'usuari pugui utilitzar els gestos que consideri més naturals, es proposa parametritzar els gestos de cada usuari mitjançant una fase d'aprenentatge en l'inici de cada sessió. Per això, per cada usuari en l'inici de la sessió, el sistema demana a l'usuari que realitzi de forma aleatòria varies execucions aïllades de cada gest. Aquesta és una forma de construir automàticament el conjunt d'entrenament. Els models construïts automàticament s'afegeixen a una base de dades, on a cada usuari se li assigna el tipus de gest que podrà realitzar. La Figura 3.12 correspon al diagrama entitat-relació de la base de dades dels models de cada usuari per cada tipus de gest. En primer lloc, l'entitat *Gest*, identifica els tipus de gestos que la interacció permet, o sigui les comandes. Per altra banda, l'entitat *Usuari* identifica cada uns dels usuaris. I la tercera entitat,

Model, emmagatzema la representació d'un gest realitzar per un usuari, o sigui el model d'uns gest. En quan a les relacions, un usuari dins el seus alfabet de gestos n'hi haurà zero o més, de la mateixa manera que un usuari tendrà zero o més models. Per altra banda, un gest podrà ser realitzat per zero o més usuaris, i tendrà zero o més models. En canvi, un model sempre farà referència a un únic gest i haurà estat realitzar per un únic usuari.

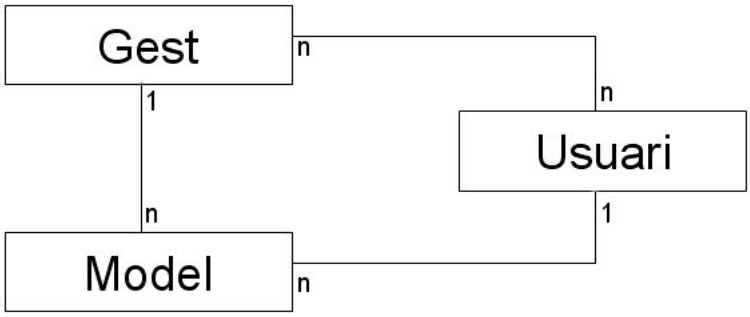

Figura 3.12: Base de dades de models de gestos

Aquesta fase d'aprenentatge permet resoldre tres problemes. En primer lloc, al demanar a l'usuari que realitzi vàries execucions aïllades de cada gest i de forma aleatòria, permet parametritzar les variacions d'estil. En segon lloc, el fet de què cada usuari tengui la seva fase d'aprenentatge, permet que aquest pugui realitzar la interacció persona-màquina amb els gestos que consideri més naturals. I finalment, el fet de que al inici de cada sessió l'usuari hagi de fer la fase d'aprenentatge permet que el sistema es recicli, encara que l'usuari ja hagués fet altres sessions. En aquest cas, a diferència de les persones que utilitzen el llenguatge de signes, que periòdicament s'han de reciclar, en aquest treball el que es recicla és el sistema en cada sessió.

Per altra banda, aquesta fase d'aprenentatge ha permès demostrar que usuaris amb el mateix origen geogràfic interpreten els gestos de forma diferent, principalment els gestos més complexes o menys habituals en la vida diària. Gestos per indicar que un objecte que apareix en la pantalla s'ha de desplaçar a la dreta o a l'esquerra són realitzats de forma bastant similar. En canvi, gestos per indicar que un objecte que apareix en la pantalla ha de rotar són interpretats de forma més diferent, aquest cas el mostra la Figura 3.13, on cada usuari interpreta l'acció a través de gestos diferents.

Figura 3.13: Interpretació del gest de la *rotacio* per diferents usuaris.

Aquest fet, permet justificar la necessitat de construir models específics dels gestos per cada usuari, amb l'objectiu de què els usuaris utilitzin els gestos que consideren més naturals a l'hora d'interaccionar.

Una vegada es sap quins gestos farà cada usuari, a partir de la fase d'aprenentatge, el sistema ha de reconèixer quan l'usuari n'està realitzant un en una sessió. Dit d'una altra manera, el sistema ha d'anar classificant el gestos que fa l'usuari en una sessió.

Existeixen vàries tècniques per dur a terme la classificació en sistemes de reconeixement de gestos. En la majoria de les propostes les propietats temporals del gest es solen resoldre estadísticament usant Hidden Markov Models (HMM), principalment degut al fet d'usar directament els valors de la imatge [76]. Així i tot, aquestes propostes no són aplicables en temps real, perquè els HMM requereixen una llarga fase d'aprenentatge per ajustar tots els paràmetres del model. La nostra idea és utilitzar el sistema de reconeixement de gestos en temps real, tenint en compte que es poden estimar les posicions 3D de les articulacions de l'usuari.

Com s'ha explicat, la representació del gest es basa en la parametrització de la seqüència de postures, i abans de començar el procés de reconeixement el sistema amb l'ajuda de l'usuari construeix un conjunt de gestos model en temps real. Aleshores, és raonable assumir que si l'usuari realitza un gest que està a prop d'algun dels gestos model, aquest gest pot ser classificat amb la classe del gest model. Per aquest motiu, s'utilitza la tècnica no paramètrica del *k-veí més proper* [22].

El *k-veí més proper* és un mètode de classificació que serveix per estimar la funció de densitat $p(x/C_j)$ de les prediccions x per cada classe C_j. És un mètode de classificació no paramètric, que estima el valor de la funció de densitat de probabilitat o directament la probabilitat *a posteriori* de que un element x pertanyi a la classe C_j a partir de la informació proporcionada per el conjunt de models. En el reconeixement de patrons, l'algorisme del *k-veí més proper* s'usa com a mètode

de classificació d'elements basat en un entrenament mitjançant exemples propers a l'espai dels elements.

Concretament, s'utilitza un classificador (k, θ) del veí més proper que trobi el k exemples de gestos més propers al gest actual que està realitzant l'usuari, i classifica aquest gest amb la classe que té el major nombre de vots, mentre tengui més de θ vots, sinó és considera que l'usuari no ha fet cap gest del conjunt model.

Per completar aquest procés, és necessari establir un quantificador que proporcioni una mesura quantitativa per comparar la similitud entre el gest actual $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ i un model $\mathbf{p} = (p_1, p_2, \ldots, p_n)$. En un primer moment es va decidir utilitzar el coeficient de Bhattacharyya [18], que és una mesura d'aproximació de la quantitat superposada entre dues mostres estadístiques. El coeficient es pot utilitzar per determinar la proximitat relativa entre dues mostres.

El càlcul del coeficient implica una forma rudimentària d'integrar la superposició de les dues mostres. Per això es pot utilitzar per comparar dos gestos si s'interpreta la representació del gest com una distribució de les variacions de la postura que ocorre quan l'usuari realitza un gest. El coeficient de Bhattacharyya ve definit de la següent manera

$$d = \sqrt{1 - \rho[\mathbf{p}, \mathbf{q}]}, \tag{3.6}$$

on

$$\rho[\mathbf{p}, \mathbf{q}] = \sum_{i=1}^{n} \sqrt{p_i q_i}, \tag{3.7}$$

és l'estimació de la mostra del coeficient de Bhattacharyya entre l'exemplar i els gest actual, i n és el nombre d'elements del vector i depèn de la representació elegida.

En el nostre cas, el coeficient de Bhattacharyya té el significat de la puntuació de la correlació entre gestos. La mesura de l'Equació 3.6 ha estat aplicada en treballs previs de seguiment visual obtenint excel·lents resultats [18].

S'ha demostrat que les mesures bin a bin, com és el cas del coeficient de Bhattacharyya, són menys robustes que les mesures que creuen els bins, ja que permeten que les característiques de diferents bins s'ajustin i permet captar la percepció de desigualtat entre distribucions. Per aquest motiu, finalment s'ha decidit utilitzar la *Earth Mover's Distance* (EMD) [60] que creua els bins i és la mesura de la quantitat de treball necessari per transformar un conjunt de punts ponderats a un altre.

3.6 Avaluació de reconeixement de gestos

Per avaluar el sistema de reconeixement de gestos, es proposa als usuaris que juguin a un videojoc interaccionant amb els gestos. Els resultats s'han estudiat en relació al temps real i al bon reconeixement dels gestos realitzats pels usuaris. Però abans de continuar amb l'avaluació, amb l'ajuda de la Figura 3.14 es presenta un esquema general del sistema. Amb el parell estèreo de càmeres es capturen imatges de l'espai de captura on es troba l'usuari. A partir d'aquestes imatges el sistema de captura dels moviments de l'usuari retorna les posicions 3D de les articulacions. Amb aquestes posicions 3D, es representa la postura de l'usuari per cada instant de temps, amb la representació proposada. A continuació es representa un gest com una representació temporal de postures. Finalment, utilitzant la representació del gest, cada gest realitzat per l'usuari és classificat per generar l'esdeveniment desitjat en temps real.

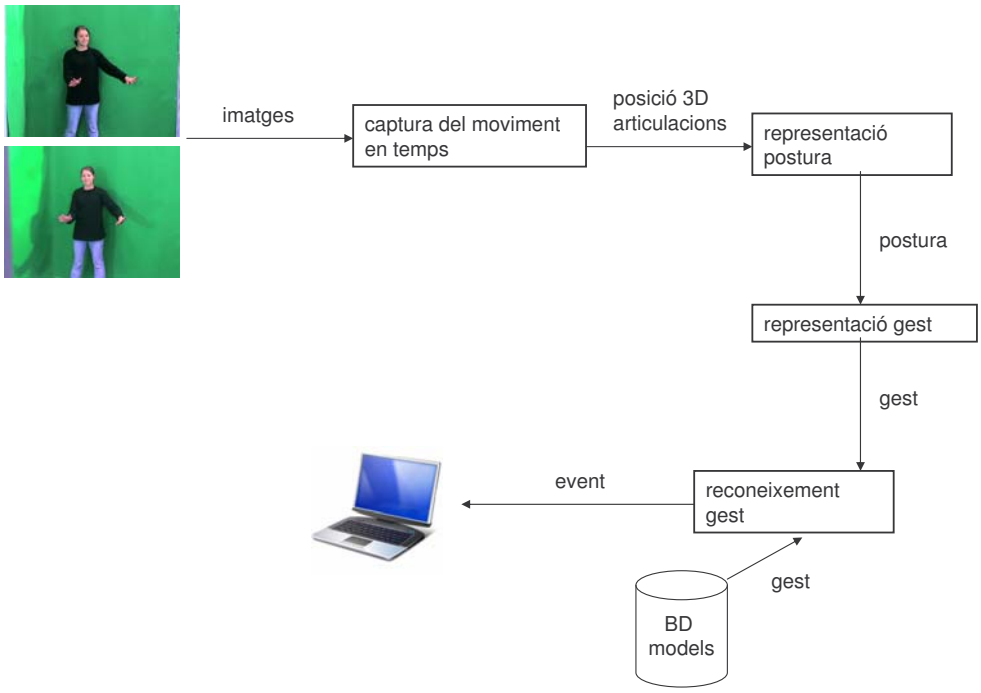

Figura 3.14: Esquema general del sistema que es presenta.

El sistema ha estat implementat amb Visual C++, utilitzant les llibreries de OpenCV [12], i s'ha provat en un context d'interacció en temps real sobre un AMD Athlon 2800 + 2.083 GHz amb Windows XP. Les imatges han estat capturades utilitzant dues càmeres DFW-500 de Sony. La resolució de la imatge que capturen les càmeres és de 320×240 i un rati de captures de 30 imatges per segon.

Una vegada repassat l'esquema general de sistema, com a prova d'avaluació de la interfície de reconeixement de gestos presentada, s'ha proposat a diferents usuaris jugar un videojoc usant gestos del cos com a forma d'interacció. En aquest cas, el joc proposat, una versió modificada del Tetris, veure Figura 3.15, permet als usuaris usar quatre formes diferents de control: *esquerra*, *dreta*, *baixar* i *rotar*.

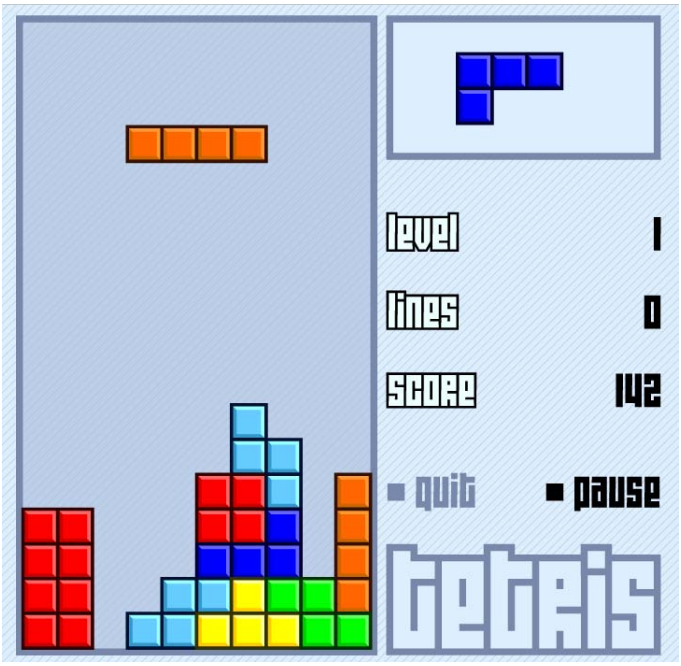

Figura 3.15: Videojoc.

El joc del Tetris, consisteix en unes peces bidimensionals de quatre blocs en diferents disposicions que cauen des de la part superior de la pantalla. El jugador no pot impedir aquesta caiguda però pot decidir el lloc on caurà la peça i la seva rotació en 0º, 90º, 180º o 270º. Quan una línia horitzontal es completa, aquesta línia desapareix i totes les peces que estan a sobre descendeixen una posició, alliberant així espai de joc i per tant facilita la tasca de situar noves peces. La caiguda dels blocs s'accelera de forma constant. El joc s'acaba quan s'amunteguen fins sortir del àrea de joc. En la Figura 3.16 es mostren les quatre formes de control del videojoc. La Figura 3.16 (a) mostra el control que permet a desplaçar la peça a l'esquerra. La Figura 3.16 (b) mostra el control que permet desplaçar la peça cap a la dreta. La Figura 3.16 (c) mostra el control que permet rotar la peça. I la Figura 3.16 (d) mostra el control que permet baixar la peça a més velocitat.

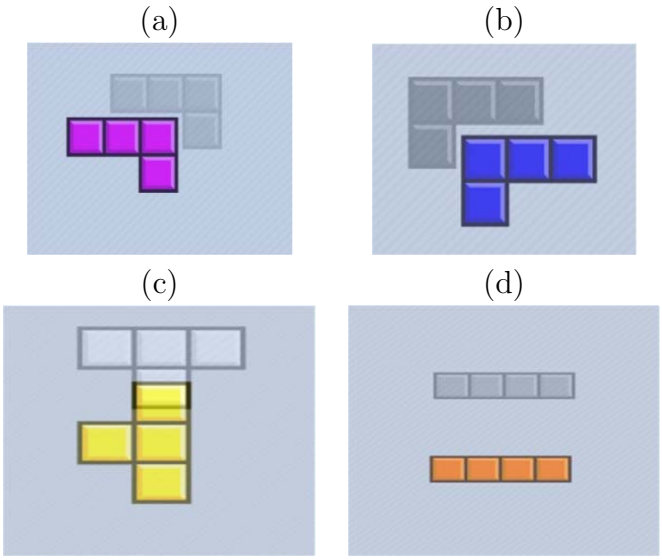

Figura 3.16: Moviments del videojoc.

Per realitzar la interacció, l'usuari s'ha de situar dins l'espai interactiu que consisteix en una pantalla de projecció i està instrumentada amb un parell estèreo de càmeres. Aquesta configuració permet a l'usuari veure el videojoc mentre realitza els gestos per controlar-lo. L'espai de captura ha de complir els requeriments que s'han explicat al Capítol 2.

A l'hora de realitzar les proves, s'han adquirit diferents sessions amb diferents usuaris que jugaven al videojoc utilitzant gestos per fer la interacció. Al mateix temps, el videojoc ha estat controlat manualment per un usuari expert perquè l'usuari tengués la sensació immersiva que realment estava jugant al joc amb els seus propis gestos. Aquest és el clàssic experiment del Mag d'Oz [33]. Aquest, és un experiment molt comú en el camp de la interacció persona-màquina, on els usuaris que interaccionen amb el sistema creuen que és autònom, quan realment és manejat per un usuari expert, però que l'usuari desconeix.

En primer lloc, per comprovar si el sistema complet treballa en temps real, s'ha calculat el temps mig que usa el sistema per tractar cada imatge, o sigui obtenir la postura de l'usuari, i a partir d'ella reconèixer els gestos que realitza. El rati és de 21 imatges per segon, per la qual cosa es pot concloure que aquesta proposta treballa prop del temps real [9].

Per altra banda, la interfície *Enactiva* ha estat avaluada per diferents usuaris que mai havien utilitzat l'aplicació, adquirint una sessió per cada usuari, en la que realitzaven les comandes necessàries per controlar el videojoc. Per tant, la base de dades d'avaluació està formada per un conjunt d'entrenament composat per diferents interpretacions de cada comanda per cada usuari diferent, i s'ha avaluat el conjunt

amb un total de 4500 frames que contenen diferents moviments de cada usuari jugant al videojoc. L'avaluació de la interfície basada en el gest ha estat realitzada amb aquest conjunt de dades. Amb l'objectiu d'interpretar millor els resultats de l'avaluació, s'han etiquetat manualment cada moviment dels usuaris de tota la base de dades amb un dels comandament de control definits prèviament. Per altra banda, en aquest experiment, hi ha dues possibles fonts d'errors: la primera, quan un moviment correcte que correspon a una comanda de control no és reconegut; la segona, quan un moviment corresponent a una comanda es reconegut com una altra comanda de control. Com a finalitat, és importat que el sistema no confongui comandes perquè podria provocar confusions d'interacció (si un usuari vol moure la peça a l'esquera i la peça es mou a la dreta). Així doncs, no és crític que el sistema no reconegui un moviment en particular (això simplement provocaria que l'usuari hauria de repetir el gest i el sistema s'estabilitzaria ell mateix). Per aquesta raó, es pot utilitzar un llindar gran per confirmar que el moviment correcte ha estat realitzat.

Els resultats obtinguts mostren que el 86% de les comandes realitzades pels usuaris, de la base de dades de seqüències, s'han reconegut correctament. El 14% de l'error és dividit en un 10.5% de moviments no reconeguts i només un 3.5% de classificacions errònies. Amb aquests resultats, s'obté un rati de reconeixement més alt que un 95% per comandes reconegudes. Des d'un punt de vista del sistema, aquest no envia cap esdeveniment al joc quan un moviment no és reconegut (només un de cada deu moviments no és reconegut), per tant es pot concloure que aquests resultats són suficients per mostrar el potencial del sistema com una interfície d'usuari fiable. A partir d'aquests resultats es pot concloure que un usuari sense preparació

pot jugar al videojoc de forma natural, només utilitzant els gestos del seu cos. A la Figura 3.17, es mostren resultats visuals del sistema.

Figura 3.17: Alguns resultats visuals del reconeixement de gestos.

Cal afegir, que la majoria de moviments no reconeguts són deguts a errors de l'estimació del sistema de captura de les articulacions del cos de l'usuari, veure Figura 3.18. Aquest fet implica que millorant el sistema de captura dels moviments de

l'usuari el reconeixement donarà millors resultats.

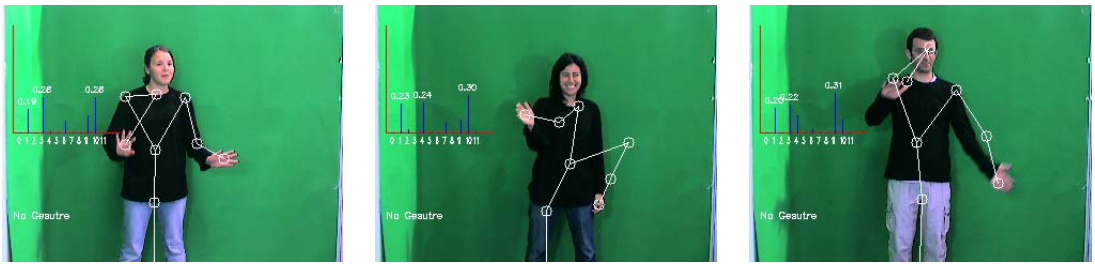

Figura 3.18: Errors de seguiment que produeixen un mal reconeixement.

A la taula Taula 3.9 es mostren els resultats separats per tipus de representació, enllaçada i acumulada. Es pot observar que la representació enllaçada és més exacte, perquè el nombre de falsos positius és més petit que en la representació acumulada, considerant un fals positiu quan el sistema reconeix un gest quan l'usuari realment no n'ha realitzat cap. A més, la representació enllaçada és més robusta en l'extracció de característiques que la representació acumulada.

Representació Postura	Correcte	Errors	No Reconeguts	Falsos Possitius
acumulada	**84.95%**	4.10%	10.95%	7.20%
enllaçada	**87.69%**	2.73%	9.58%	4.18%

Taula 3.9: Resultats comparatius entre les representacions proposades de la postura.

3.7 Resum

En aquest capítol s'ha presentat un sistema de reconeixement de gestos. El sistema resol els principals problemes del reconeixement de gestos: les variacions espacials,

les variacions temporals i les variacions d'estil. En primer lloc, les variacions espacials es resolen amb una representació de la postura, que permet generalitzar sobre les diferències de la forma del cos en la població dels usuaris. A continuació, les variacions temporals es resolen utilitzant una representació temporal del gest, ja que es considera que un gest està format per una seqüència de postures. Finalment, les variacions d'estil es resolen, parametritzant els gestos de cada usuari mitjançant una fase d'aprenentatge en el començament de cada sessió.

S'ha mostrat el potencial del sistema des d'una interfície *Enactiva*, que consistia en el control d'un videojoc en temps real a través de gestos. Els resultats mostren, que el sistema presentat obté un rati de reconeixement superior al 95% per a moviments reconeguts.

Capítol 4

Restricció basada en la imatge per a la cinemàtica inversa

Quan la llei de les matemàtiques és refereix a la realitat,
aquesta no és exacte;
i quan les matemàtiques són exactes,
aquestes no es refereixen a la realitat.
Albert Einstein.

En el Capítol 3, s'ha vist que alguns resultats del sistema de captura no eren acceptables, cosa que provocava que el reconeixement de gestos fos erroni. Per aquest motiu, en aquest capítol es presenta com afegir una restricció basada en la imatge per millorar els resultats del sistema de captura del moviment humà.

4.1 Enfocament

El sistema de captura de moviment de l'usuari es basa en la combinació d'algorismes de visió per ordinador i de cinemàtica inversa, i tendeix a simplificar la proposta combinant els resultats d'ambdues tècniques que són aplicades per separat. Com mostra la Figura 4.1, la proposta actual es basa en la detecció de certes articulacions en la imatge usant algorismes de visió per ordinador. Aleshores, utilitzant cinemàtica inversa, s'estima una postura del cos plausible. D'aquesta forma, l'algorisme de cinemàtica inversa és *cec* respecte la projecció del segment del cos dins les imatges. El sistema de captura només utilitza restriccions biomecàniques per estimar la millor posició de les articulacions no detectades en les imatges.

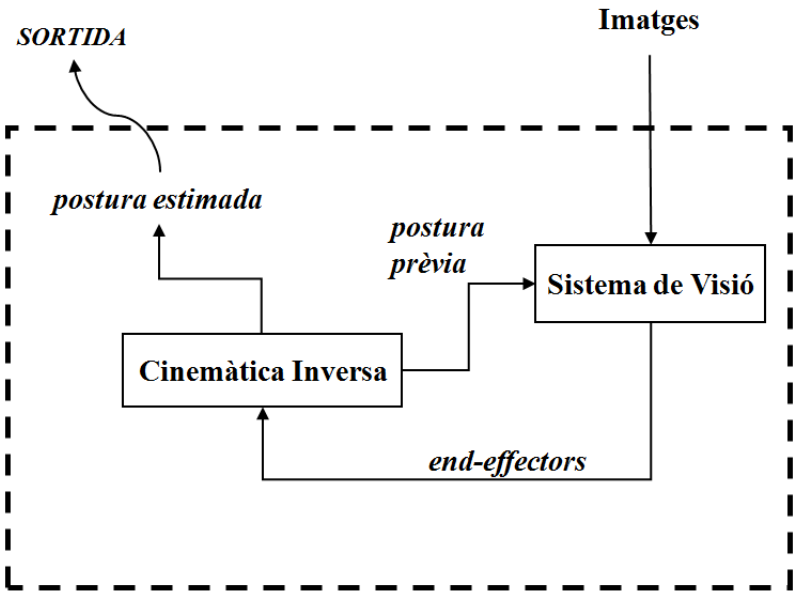

Figura 4.1: Arquitectura general del sistema.

Les imatges que es mostren a la Figura 4.2, són un exemple clar d'aquest fet, on la cinemàtica inversa és cega a l'hora d'estimar les posicions de les articulacions que no són els *end-effectors*, ja que només treballa amb restriccions biomecàniques.

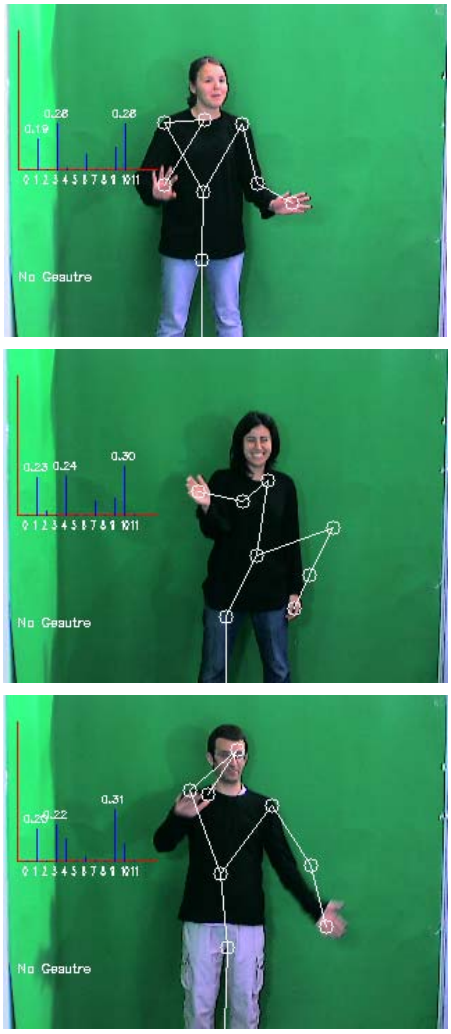

Figura 4.2: Errors de tracking que produeixen un mal reconeixement.

En aquest capítol es presenta una nova proposta on l'objectiu és incloure informació de la imatge directament dins l'esquema de la cinemàtica inversa, veure Figura 4.3. La idea és afegir, a l'esquema anterior, restriccions d'imatge per limitar la redundància de solucions de les cadenes cinemàtiques. Addicionalment, també es vol donar a conèixer que és possible utilitzar imatges preprocessades, en altres paraules, l'algorisme de visió per ordinador podria processar les imatges d'entrada amb l'objectiu de fer el problema més tractable o per millorar les característiques desitjades de la imatge per aplicacions específiques. Finalment, es vol mostrar que aquest esquema per reconstruir la postura pot ser usat amb una o més vistes, això vol dir que pot treballar només amb una vista però que els resultats milloren si s'afegeixen més vistes de l'escena.

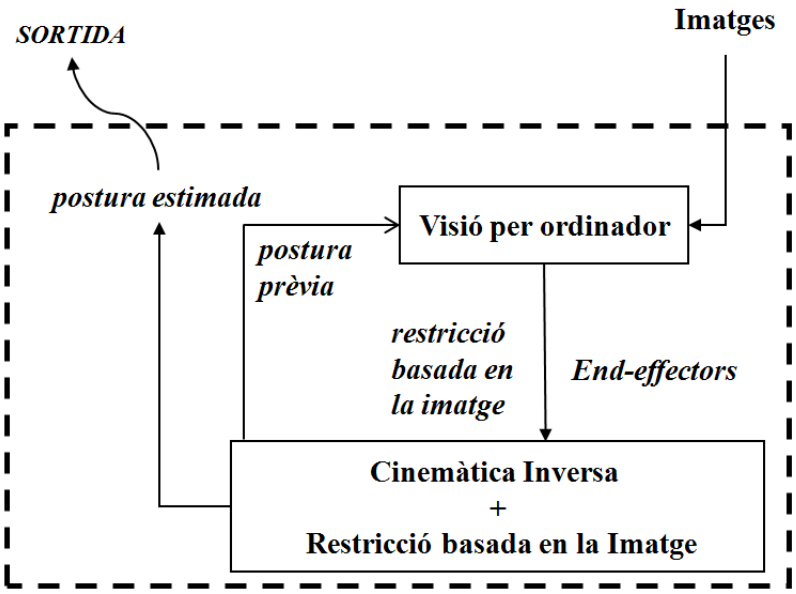

Figura 4.3: Arquitectura general del sistema amb restricció basada en imatge per la cinemàtica inversa.

Amb l'objectiu d'avaluar aquesta proposta, s'han dut a terme diferents experiments. Primerament, s'han utilitzat imatges sintètiques per mostrar que la proposta funciona en una situació ideal. Aquest cas simple mostra com de forma teòrica la proposta funciona. A continuació, per avaluar el seu potencial en situacions reals, es mostren els resultats d'experiments amb seqüències reals. En aquests experiments s'usen una seqüència anotada, ja utilitzada en el Capítol 2, i una coneguda base de dades de moviments humans que conté la informació de captura del moviment. Aquests experiments permeten estudiar quantitativament les diferents seqüències amb l'objectiu d'avaluar el funcionament d'aquesta proposta.

4.2 Treballs previs

En aplicacions biomecàniques on l'objectiu és estudiar el moviment humà, una component crítica és obtenir amb exactitud la posició 3D de les articulacions del cos de l'usuari. Normalment, els mètodes més comuns per obtenir les posicions de les articulacions requereixen un ambient de laboratori i l'ús de marcadors sobre el cos de l'usuari. Aplicacions modernes de biomecànica requereixen una captura exacte dels moviments humans habituals sense la utilització d'artefactes associats amb els sistema de captura estàndards basats en marcadors [41]. Noves tècniques i la recerca en visió per ordinador ha provocat un ràpid desenvolupament d'enfocaments de captura del moviment sense marcadors.

En visió per computador, els algorismes són dissenyats per permetre al sistema analitzar una o múltiples seqüències d'imatges amb l'objectiu de recuperar el movi-

ment humà. El problema és que les imatges són en 2D i els moviments humans en 3D. Aquest fet provoca que es produeixin ambigüitats; hi ha un gran nombre de configuracions 3D del cos humà que podrien ser explicades amb una imatge. Endemés, les imatges poden contenir renou i interferències, inclús estar incompletes (algunes articulacions o segments poden no ser visibles). Aleshores, només es pot realitzar una estimació de la postura de l'usuari. La cinemàtica inversa, com hem vist, pot ajudar a realitzar aquesta estimació. Per exemple, en el treball de Zhou [80], els angles de les articulacions són estimats utilitzant cinemàtica inversa basada en restriccions de l'esquelet, i les coordenades dels píxels en els segments del cos en l'escena són determinats per cinemàtica directe. Finalment, la postura del moviment humà pot ser reconstruïda mitjançant histogrames. L'inconvenient d'aquesta proposta és que l'algorisme no permet moviments humans en la direcció de desplaçament Z. En el cas d'utilitzar múltiples càmeres, les ambigüitats es poden reduir, com s'ha vist en el Capítol 2.

4.3 Restricció basada en la imatge

Com s'ha explicat al Capítol 2, és possible restringir les solucions de la cinemàtica inversa afegint un criteri escalar $h(\boldsymbol{\theta})$, com es mostra a la següent equació.

$$\Delta\boldsymbol{\theta} = J^{+\lambda}\Delta\mathbf{x} - \alpha P_{N(J)}\nabla h(\boldsymbol{\theta}), \tag{4.1}$$

on α és un factor de guany positiu que depèn de la configuració. La definició de la tasca secundaria a través del criteri $\nabla h(\boldsymbol{\theta})$ depèn de l'aplicació. Per definició,

l'espai nul del jacobià $N(J)$ s'assigna per J en el vector nul de l'espai restringit de variacions. Dit més planerament, la variació del vector a través de $N(J)$ no té efectes sobre les restriccions. L'equació 4.2 correspon a la projecció de l'espai nul.

$$P_{N(J)} = I_n - J^+ J, \tag{4.2}$$

on I_n és la matriu identitat $n \times n$.

A continuació, s'explica com definir aquest criteri utilitzant imatges, amb l'objectiu de guiar la reconstrucció de la postura de la cadena articulada per les aplicacions de captura del moviment humà.

Per definir la restricció basada en la imatge, aquesta proposta es basa en el treball de Servo Control Visual [16]. Concretament, en el treball de Marchand i Courty, on defineixen una tasca secundària basada en la imatge per controlar una càmera en entorns virtuals [40]. Per propòsits de captura del moviment, s'ha de tenir en compte que l'estructura de model humà és altament redundant, per tant existeix un ampli espai de solucions. Com s'ha vist al Capítol 2, una solució per limitar l'espai de solucions és incloent més tasques, usant una estratègia de prioritats [2]. En aquest cas, la solució garanteix que una tasca associada amb una prioritat alta serà assolida tan com sigui possible, mentre que una restricció amb prioritat baixa serà optimitzada només en el reduït espai de solucions que no influenciï les tasques de prioritat major. A l'hora d'explicar la proposta de la restricció basada en la imatge (o ibIK), es farà amb un esquema de dues tasques perquè sigui més entenedor, o sigui que l'explicació es basarà en l'Equació 4.1. Estendre-ho a més prioritats és fàcil si la restricció de la imatge té la prioritat més baixa. Endemés, és possible utilitzar el mètode del Jacobià

estès [34] amb l'objectiu de donar la prioritat més alta a la restricció basada en la imatge.

Per aplicacions de captura del moviment, es defineix $h(\boldsymbol{\theta})$ amb l'objectiu de maximitzar la coincidència entre la projecció de la cadena cinemàtica dins les imatges i el cos humà. Consideris el cas de la Figura 4.4(a), on es mostra la configuració inicial de la cadena cinemàtica i l'objectiu és estimar la posició del colze a partir de la posició 3D de la mà com *end-effector*. Aplicant Cinemàtica Inversa, s'obté el resultat de la Figura 4.4(b), on l'estimació del colze recau fora del cos degut a la natura *cega* de només usar les posicions desitjades dels *end-effectors*. Amb l'objectiu de resoldre aquest problema, es proposa un criteri que provi de guiar la cadena cinemàtica a la projecció del cos de les imatges, Figura 4.4(c). Es vol ressenyar, que encara que l'articulació de la cadena cinemàtica que representa l'espatlla de l'usuari, pareixi que està situada fora de la seva silueta, aquesta està situada correctament. El que succeeix és que al dibuixar la cadena cinemàtica, s'ha elegit una gruixa de línia que fa que el centre de l'articulació també s'hagi eixamplat.

(a)

(b)

(c)

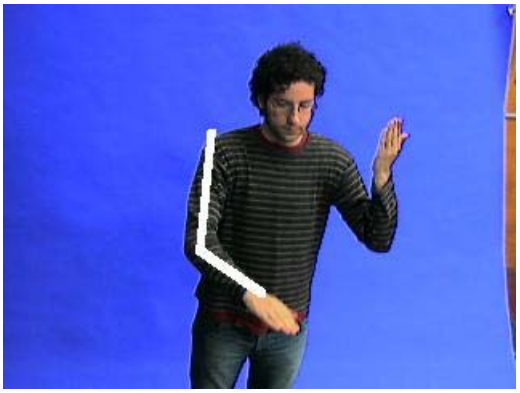

Figura 4.4: (a) Configuració inicial de la cadena cinemàtica. (b) Estimació IK. (c) Resultat del IK utilitzant una restricció basada en la imatge.

Formalment, es defineix $h(\boldsymbol{\theta})$ de la següent manera

$$h(\boldsymbol{\theta}) = \frac{1}{n} \sum_{c} \sum_{x} \sum_{y} \left(\boldsymbol{I}_c(x, y) \cdot \boldsymbol{M}_c(x, y, \boldsymbol{\theta}) \right), \tag{4.3}$$

on $\boldsymbol{I}_c(x, y)$ representa la intensitat del punt 2D (x, y) de la imatge c ($c \geq 1$), que correspon a diferents vistes de l'usuari, i n és el nombre de punts (x, y) que pertanyen a la imatge de suport desitjada $\boldsymbol{I}_c$. $\boldsymbol{M}_c(x, y, \boldsymbol{\theta})$ serà explicada més endavant. Aplicant un algorisme d'eliminació de fons [32], és possible usar directament la silueta $\boldsymbol{I}$ com la imatge de suport, de la cadena cinemàtica, veure Figura 4.5(a). Amb l'objectiu d'obtenir una superfície més suau s'aplica la transformació de la distància Euclídia [3] a la imatge de la silueta, veure Figura 4.5(b). Ambdues operacions són ràpides i no introdueixen cap retràs a l'algorisme.

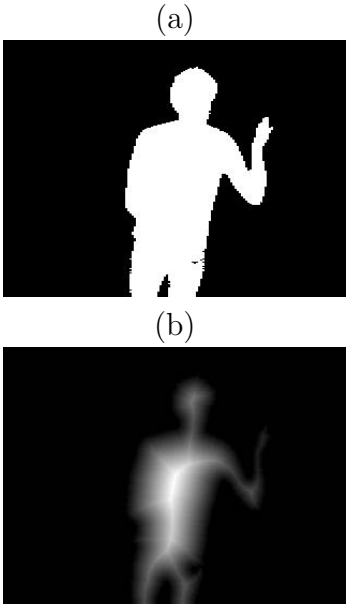

Figura 4.5: Imatge de suport: (a) silueta; (b) transformació de la distància Euclídia.

Amb la finalitat de completar la definició de l'Equació 4.3, a continuació es defineix la funció $\boldsymbol{M}_c(x, y, \boldsymbol{\theta})$ on (x, y) és un punt de la projecció de la cadena respecte a la càmera c i $\boldsymbol{\theta}$ és la configuració de la cadena cinemàtica. Si $\boldsymbol{X} = (X, Y, Z)$ són les coordenades de l'articulació i en l'espai 3D, i s'assumeix que la informació del calibratge és coneguda amb l'objectiu de projectar les coordenades 3D en les imatges 2D, es defineix $\boldsymbol{p}_{c,i} = (x, y)$ com les coordenades 2D de la imatge de la projecció de l'articulació i dins la imatge c. Assumint que les articulacions estan ordenades de forma consecutiva, la funció $\boldsymbol{M}_c(x, y, \boldsymbol{\theta})$ és definida de la següent forma

$$\boldsymbol{M}_c(x, y, \boldsymbol{\theta}) = \begin{cases} 1, & \text{if } (x, y) \in \overline{\boldsymbol{p}_{c,i}\boldsymbol{p}_{c,i+1}} \text{ per una articulació } i \\ 0, & \text{en cas contrari} \end{cases} \tag{4.4}$$

on $\overline{\boldsymbol{p}_{c,i}\boldsymbol{p}_{c,i+1}}$ és el segment entre la projecció de les articulacions 3D en la imatge. La Figura 4.6 mostra la funció $\boldsymbol{M}_c(x, y, \boldsymbol{\theta})$ per l'exemple de la Figura 4.4(a) i les seves derivades parcials. Per tant, la restricció basada en la imatge ve donada pel gradient del criteri $h(\boldsymbol{\theta})$, que es mostra en la següent equació.

$$\nabla h(\boldsymbol{\theta}) = \left(\frac{\partial h(\boldsymbol{\theta})}{\partial \theta_j} \right), \ j = 1, \ldots, n, \tag{4.5}$$

on la derivada parcial de l'articulació j ve donada per

$$\frac{\partial h(\boldsymbol{\theta})}{\partial \theta_j} = \frac{1}{n} \sum_c \sum_x \sum_y \left(\boldsymbol{I}_c(x, y) \cdot \frac{\partial \boldsymbol{M}_c(x, y, \boldsymbol{\theta})}{\partial \theta_j} \right). \tag{4.6}$$

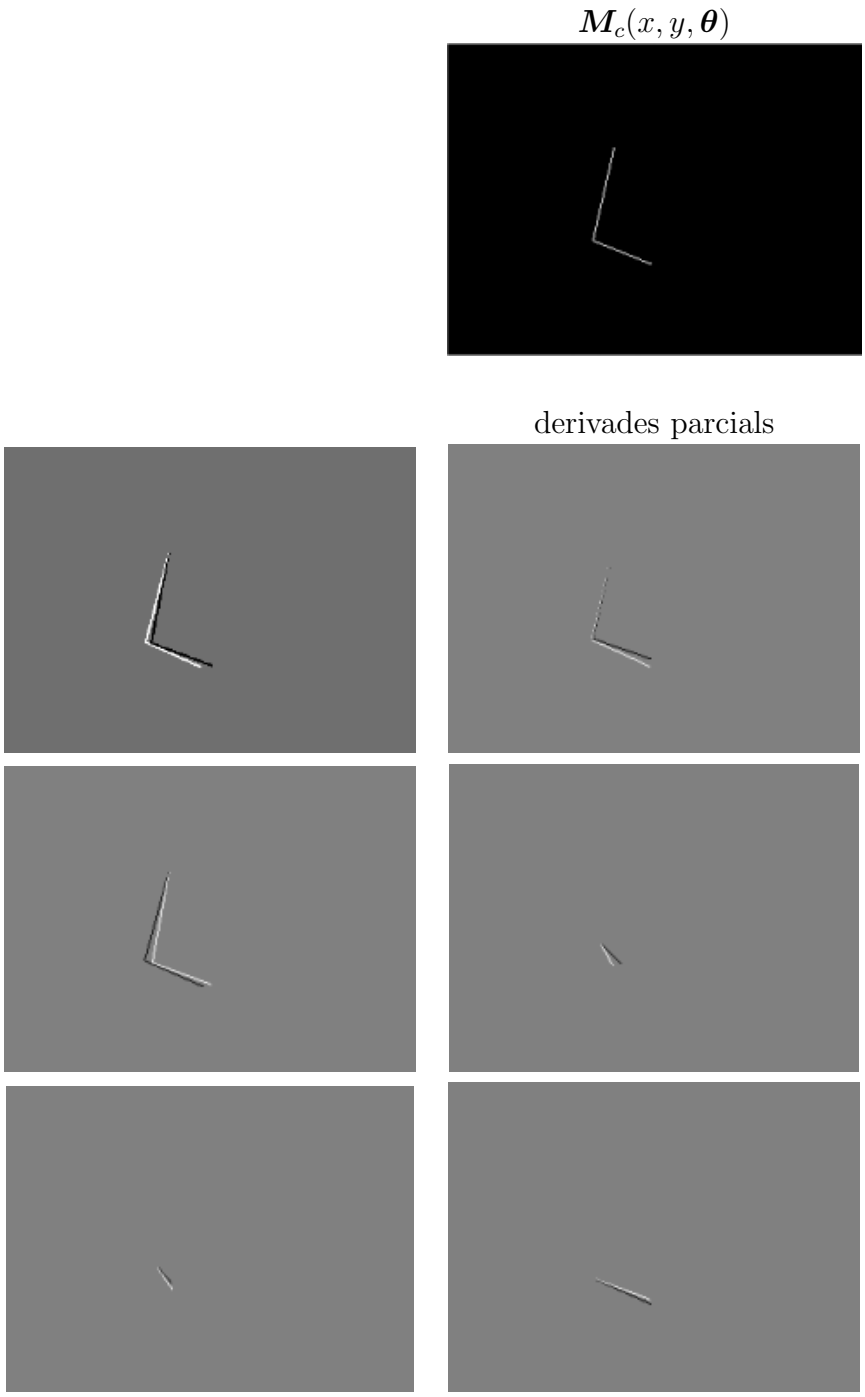

Figura 4.6: La funció $\boldsymbol{M}_c(x, y, \boldsymbol{\theta})$ i les seves derivades parcial en el cas de la Figura 4.4(a).

4.4 Avaluació

L'enfocament proposat en aquest capítol s'ha avaluat amb quatre proves diferents. La primera prova utilitza un entorn virtual per demostrar que l'enfocament presentat funciona correctament en una situació ideal. La segona prova aplica l'enfocament presentat en aquest capítol sobre un conjunt de seqüències de moviments de l'usuari per mostrar que funciona bé amb imatges reals. Per altra banda, en el tercer experiment es compara l'avaluació de la cinemàtica inversa amb la restricció basada en la imatge i sense, utilitzant el conjunt de dades de l'HumanEva [61]. Aquest conjunt de dades està format per quatre subjectes que realitzen sis tipus d'accions diferents capturades en set seqüències de vídeo calibrades des de diferent punts de vista. Cal destacar, que les seqüències de vídeo estan sincronitzades amb els seus corresponents paràmetres 3D de la postura dels moviments capturats.

El sistema ha estat implementat amb Visual C++, utilitzant les llibreries de OpenCV [12], i s'ha provat en un context d'interacció en temps real sobre un AMD Athlon 2800 + 2.083 GHz amb Windows XP.

4.4.1 Entorn virtual

En primer lloc, es prova el sistema en un entorn virtual per demostrar que l'entorn funciona correctament en una situació ideal. Per fer-ho, s'ha definit una cadena cinemàtica en l'espai 2D, amb 4 articulacions rotacionals d'un grau de llibertat cada una. Per avaluar el sistema, aleatòriament es genera una configuració inicial i una configuració objectiu de la cadena cinemàtica. A continuació, s'aplica l'enfocament

de Cinemàtica Inversa, des de la configuració inicial amb la restricció basada en la imatge i sense, per estimar la configuració objectiva de la cadena cinemàtica, amb l'extrem de la configuració objectiu com *end-effector*.

En aquest primer experiment, que es pot visualitzar a la Figura 4.7, es pot observar que si no s'utilitza la restricció basada en la imatge, quan la cadena cinemàtica assoleix la posició de l'*end-effector* l'estimació de la configuració objectiu s'atura. En canvi, si s'utilitza la restricció basada en la imatge aquesta continua intentant inserir la cadena cinemàtica dins la projecció de la configuració objectiu, encara que la cadena cinemàtica hagi arribat a la posició de l'*end-effector*.

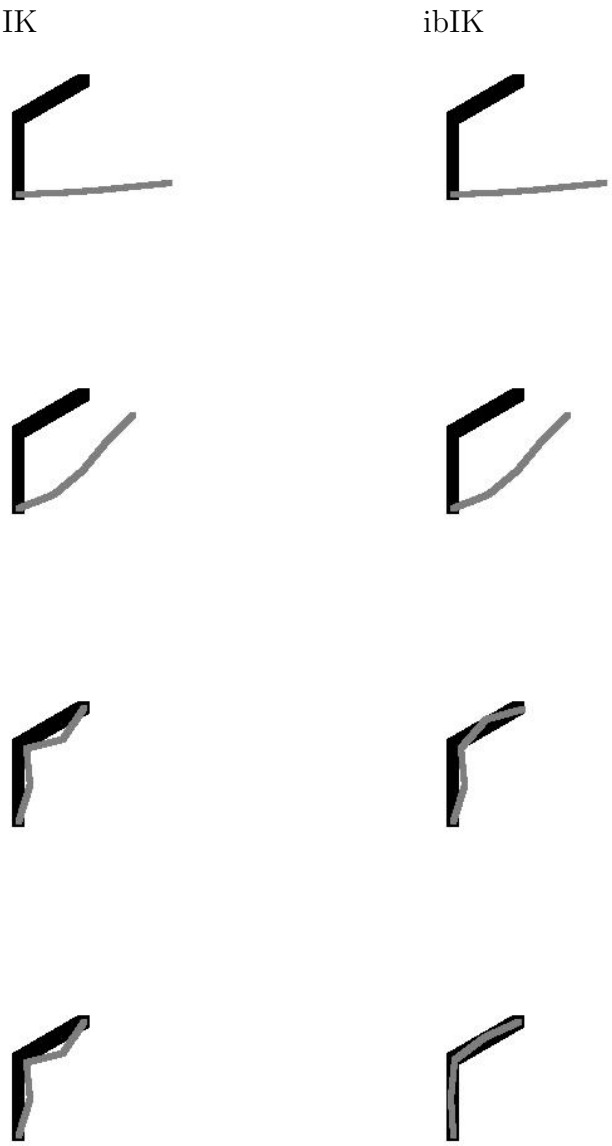

Figura 4.7: Experiment 1: A partir de la configuració inicial de la cadena cinemàtica (gris), ambdós enfocaments proven d'assolir la configuració objectiu (negre). Columna esquerra: seqüència on s'aplica la Cinemàtica Inversa sense utilitzar la restricció basada en la imatge. Columna dreta: seqüència on s'aplica la Cinemàtica Inversa utilitzant la restricció basada en la imatge.

En la Figura 4.8 es mostra un segon experiment en l'entorn virtual, on aleatòriament es torna generar una configuració inicial i una objectiu de la cadena cinemàtica. Però en aquest cas ambdues configuracions comparteixen la posició de l'extrem, o sigui el mateix *end-effector*. Es torna realitzar el mateix experiment, des de la configuració inicial s'intenta estimar la configuració objectiu amb cinemàtica inversa, utilitzant la restricció basada en la imatge i sense utilitzar-la. Els resultats mostren que si no s'utilitza la restricció basada en la imatge la configuració inicial no canvia, ja que l'*end-effector* s'ha assolit des d'un primer moment. En canvi, si s'utilitza la restricció basada en la imatge s'obliga a la configuració inicial assolir també la projecció de la configuració objectiu.

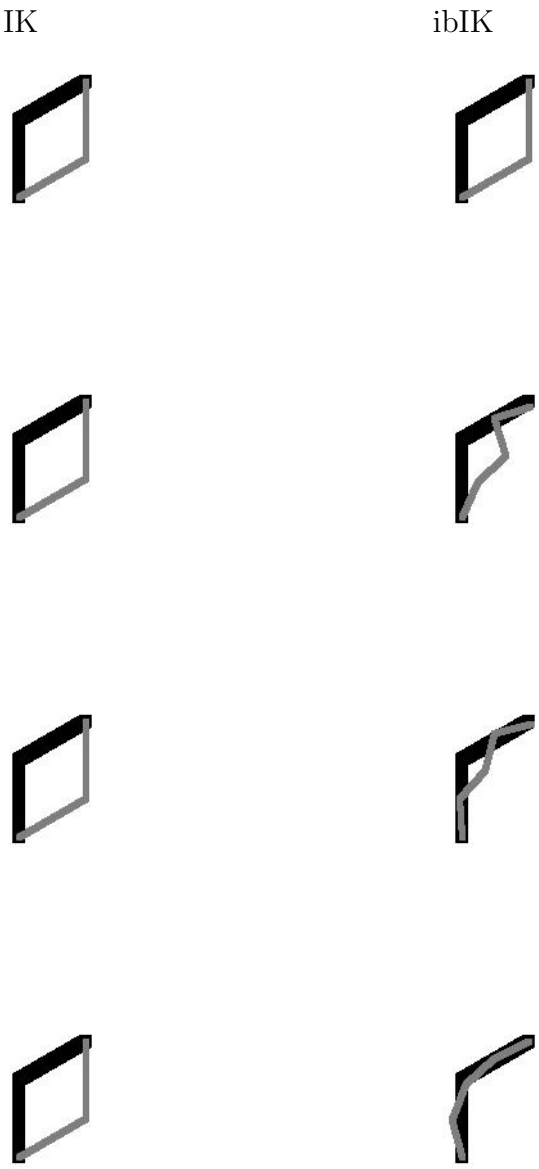

Figura 4.8: Experiment 2: Configuració inicial (gris) i objectiu (negre) amb el mateix *end-effector*. Columna esquerra: seqüència on s'aplica la Cinemàtica Inversa sense utilitzar la restricció basada en la imatge. Columna dreta: seqüència on s'aplica la Cinemàtica Inversa utilitzant la restricció basada en la imatge.

El darrer experiment en l'entorn virtual es mostra a la Figura 4.9, on la cadena cinemàtica intenta evitar un objecte projectat en la imatge de suport. Aleatòriament es defineix una configuració inicial de la cadena cinemàtica, la posició de l'*end-effector* objectiu, i un objecte rectangular. En aquest cas, s'utilitza la restricció de la imatge per evitar l'objecte. Els resultats mostren que afegint la restricció basada en la imatge ajuda a la cinemàtica inversa, en el sentit que li permet assolir l'*end-effector* i evitar l'objecte rectangular.

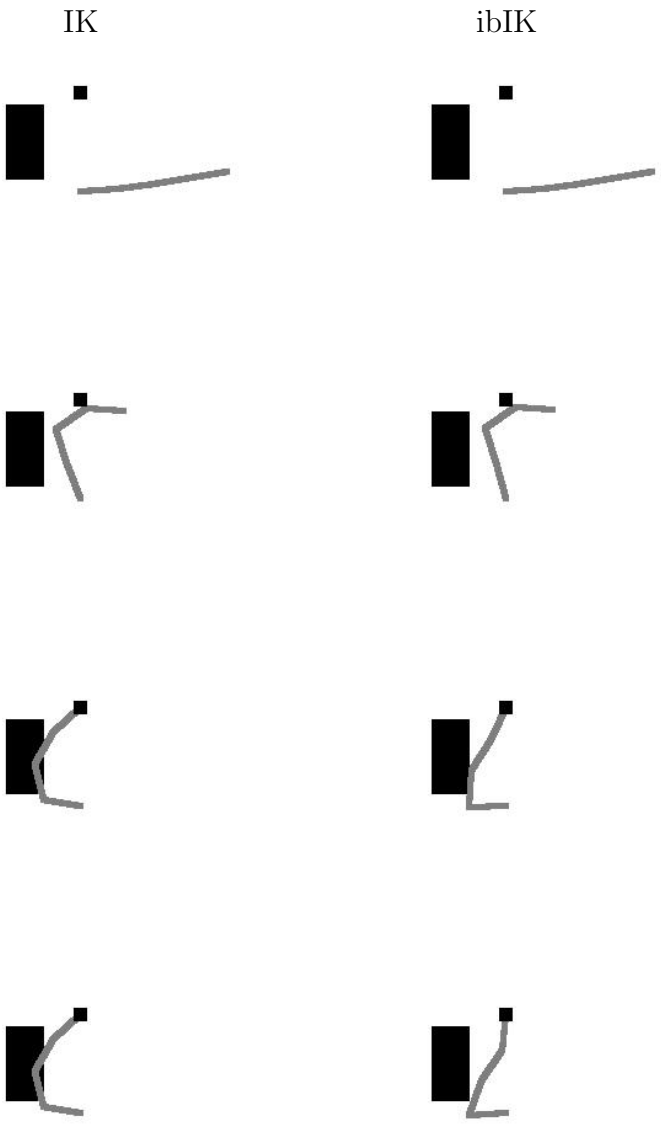

Figura 4.9: Experiment 3: A partir de la configuració inicial (gris) de la cadena cinemàtica, l'experiment intenta evitar un objecte (negre). Columna esquerra: seqüència on s'aplica la Cinemàtica Inversa sense utilitzar la restricció basada en la imatge. Columna dreta: seqüència on s'aplica la Cinemàtica Inversa utilitzant la restricció basada en la imatge.

4.4.2 Imatges reals

En aquest test, s'aplica la cinemàtica inversa amb la restricció basada en la imatge sobre una seqüència estereoscòpica real de moviments d'una persona. A més, les posicions 3D de les articulacions de l'usuari que apareix en la seqüència estan anotades de forma manual per a poder realitzar una comparació quantitativa. La seqüència està formada per 450 imatges estèreo que corresponen a 15 segons en temps real. L'objectiu principal d'aquesta prova és mostrar que l'enfocament proposat funciona bé en imatges reals. Endemés, aquest experiment mostra com aquest enfocament és capaç de solucionar el problema utilitzant només una imatge.

En la Figura 4.10 es mostra una imatge de la seqüència estereoscòpica aplicant cinemàtica inversa utilitzant la restricció basada en la imatge i sense utilitzar-la. Per fer-ho, es defineix una cadena cinemàtica en l'espai 3D, amb 2 articulacions rotacionals de 3 graus de llibertat cada una. Aquesta cadena cinemàtica modela el braç de l'usuari que apareix en la seqüència. En el cas en què s'utilitza la restricció basada en la imatge, s'aplica primer l'algorisme usant només la càmera esquerra ($c = 1$), i després usant les dues càmeres ($c = 2$). Els resultats mostren que quan la cinemàtica inversa no utilitza la restricció basada en la imatge, aquest perd la posició del colze. En canvi, la cinemàtica inversa utilitzant la restricció basada en la imatge estima la posició del colze dins la silueta de l'usuari. En aquest cas, utilitzant seqüències estereoscòpiques, no hi ha diferencies significants entre usar una imatge o usar les dues.

Figura 4.10: Imatges reals. Primera fila: configuració inicial de la cadena cinemàtica. Segona fila: Estimació del braç aplicant cinemàtica inversa. Tercera fila: Estimació del braç aplicant cinemàtica inversa utilitzant la restricció basada en la imatge amb una vista. Quarta fila: Estimació del braç aplicant cinemàtica inversa utilitzant la restricció basada en la imatge amb dues vistes.

Amb l'objectiu de realitzar una avaluació quantitativa s'utilitza l'error quadràtic (mitjà). Formalment, l'error entre una posició 3D estimada $\boldsymbol{X^e}$ i la que realment es fa $\boldsymbol{X^{GT}}$, que es considera la real, es calcula com

$$E(\boldsymbol{X}^e, \boldsymbol{X}^{GT}) = \frac{1}{i} \sum ||\boldsymbol{X}^e - \boldsymbol{X}^{GT}||_2 \tag{4.7}$$

on i és el nombre de punts i $|| \cdot ||_2$ és la norma euclidiana. Concretament, en aquest experiment es compara les posicions dels colzes anotades manualment i la posicions dels colzes estimades amb la cinemàtica inversa utilitzant la restricció basada en la imatge en els cas d'una i dues vistes. Cal afegir, que també es compara amb els resultats de la cinemàtica inversa prioritzada (PIK) mesurats a [9], ja que la mateix prova es va utilitzar per avaluar el seu funcionament. La Taula 4.1 resumeix els resultats que mostren que la cinemàtica inversa quan utilitza la restricció basada en la imatge té un error menor a l'hora d'estimar la posició del colze. També es pot observar, com en aquest cas, utilitzar una segona vista no millora significativament els resultats.

	PIK (mm)	ibIK-1 vista (mm)	ibIK-2 vistes (mm)
colze esquerra	46.54	20.05	19.81
colze dret	42.40	19.86	19.07

Taula 4.1: Comparació amb la seqüència anotada manualment. PIK: Cinemàtica inversa prioritzada. ibIK: cinemàtica inversa utilitzant la restricció basada en la imatge.

4.4.3 HumanEva

En aquesta prova s'avalua el sistema presentat en aquest capítol, utilitzant dues vistes de dues seqüències reals de moviments, *walking* i *box*, del subjecte 1 de l'HumanEva. Aquestes seqüències estan formades per un total de 3050 imatges per vista. Pel fet que aquesta base de dades també disposa de les posicions 3D de les articulacions gràcies a uns marcadors i a un sistema de captura, l'objectiu dels següents experiments és realitzar una avaluació quantitativa de l'enfocament presentat en aquest capítol. Per dur-ho a terme, s'ha definit una cadena cinemàtica en l'espai 3D, amb dues articulacions rotacionals de tres graus de llibertat cada una, amb l'objectiu de modelar i estimar la configuració del braç (en la seqüència *box*) i la cama en la seqüència (*walking*).

Utilitzant l'error l'error quadràtic (mitjà) de l'Equació 4.7, la Taula 4.2 mostra els resultats obtinguts sobre les dues seqüències aplicant la cinemàtica inversa utilitzant la restricció basada en la imatge i sense utilitzar-la

	IK (mm)	ibIK (mm)
seqüència *box*	47.72	21.39
seqüència *walking*	40.69	16.35

Taula 4.2: Error global de l'estimació de les posicions 3D de l'articulació interna de les dues seqüències BOX i WALKING (el colze en el cas del braç i el genoll en el cas de la cama). IK: cinemàtica inversa. ibIK: cinemàtica inversa amb restricció basada en la imatge.

Els resultats visuals per la seqüència *box* es mostren en les Figures 4.11 i 4.12, on es pot veure que afegint la restricció basada en la imatge, l'estimació del colze recau dins la silueta. En canvi si no s'usa la posició del colze es situa fora.

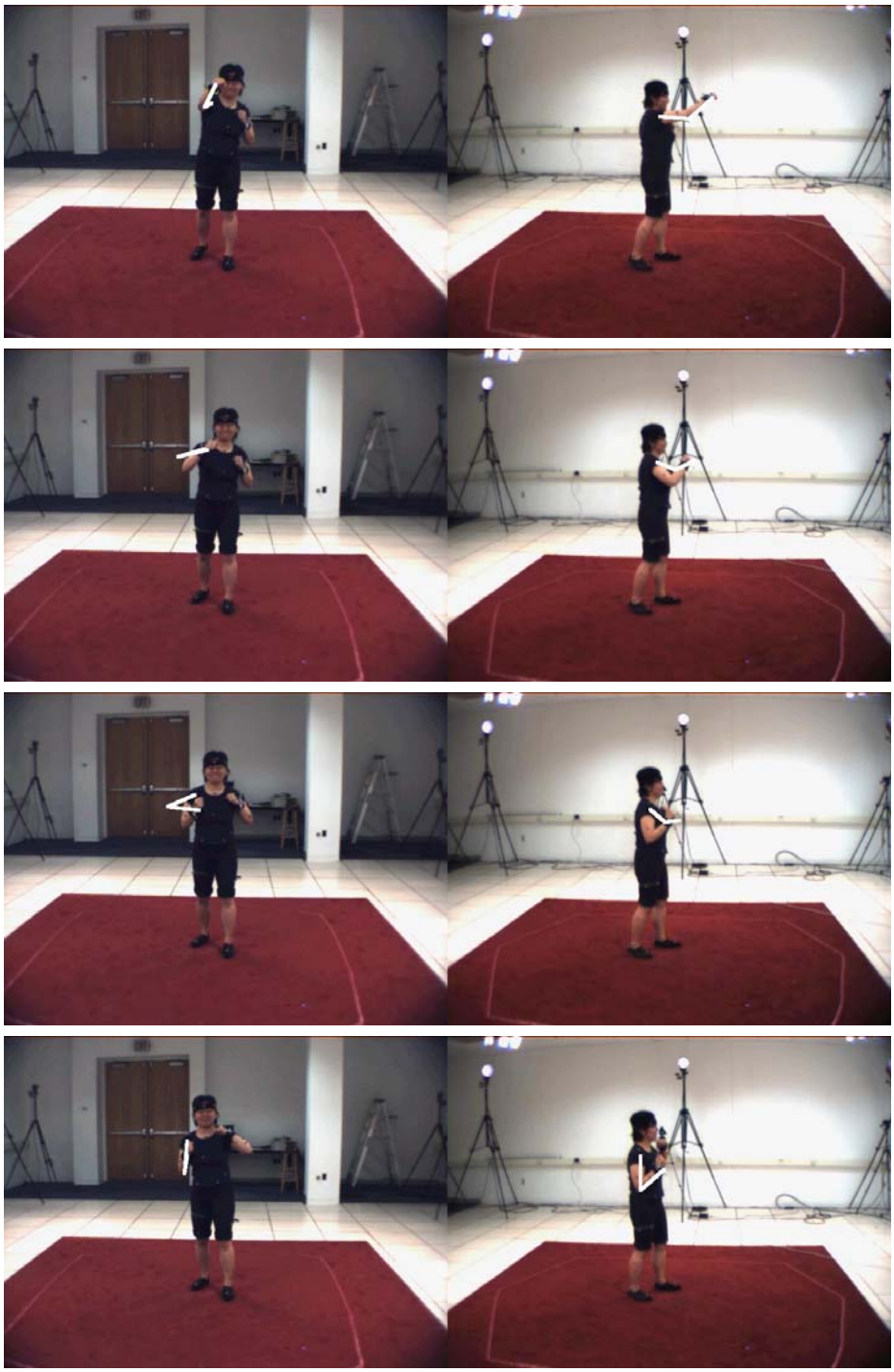

Figura 4.11: Seqüència *box* de l'HumanEva per el subjecte 1 aplicant IK.

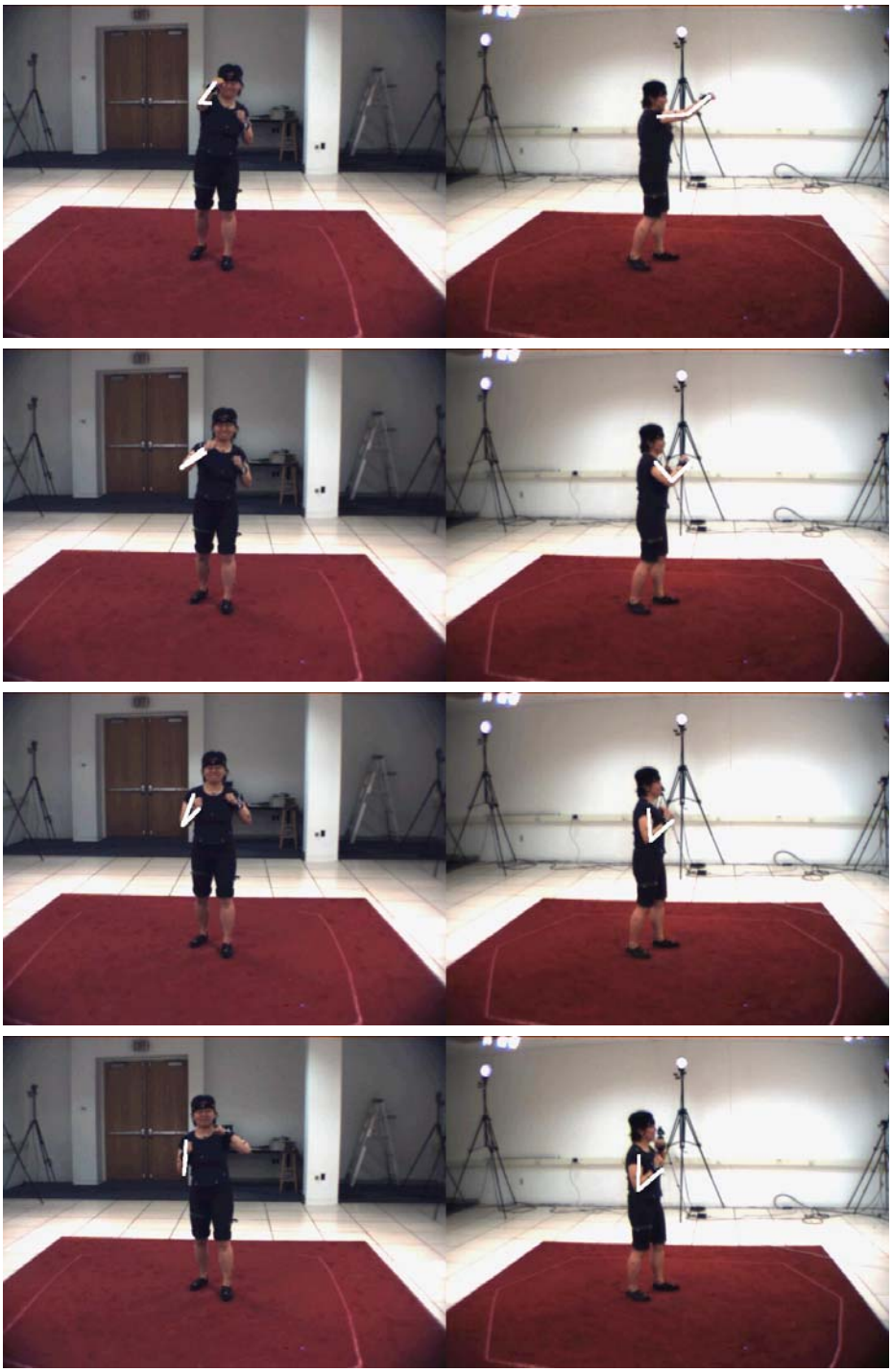

Figura 4.12: Seqüència *box* de l'HumanEva per el subjecte 1 aplicant ibIK.

En el gràfic de la Figura 4.13, per la seqüència *box*, es mostra l'error de l'estimació del colze al aplicar cinemàtica inversa sense l'ajuda de la restricció basada en la imatge i es compara amb l'error de l'estimació de la posició del colze al aplicar cinemàtica inversa amb una restricció basada en la imatge. Es pot observar que en la majoria dels casos l'estimació és millor amb ibIK.

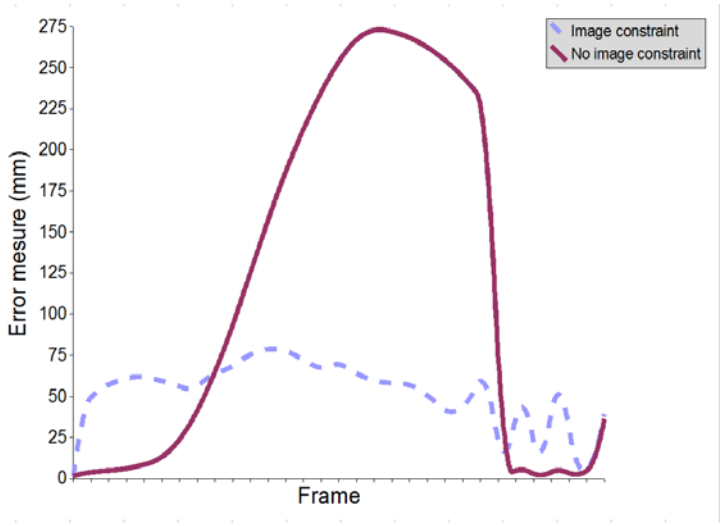

Figura 4.13: Estimació del colze per cada imatge de la seqüència *box*.

Les Figures 4.14 i 4.15 mostren els resultats d'aplicar la proposta IK i la proposta ibIK a la seqüència *walking*. En aquest cas, si s'utilitza la restricció basada en la imatge el moviment de la cama recuperat és més natural, i evita les irregularitats causades per la cinemàtica inversa en l'estimació del genoll.

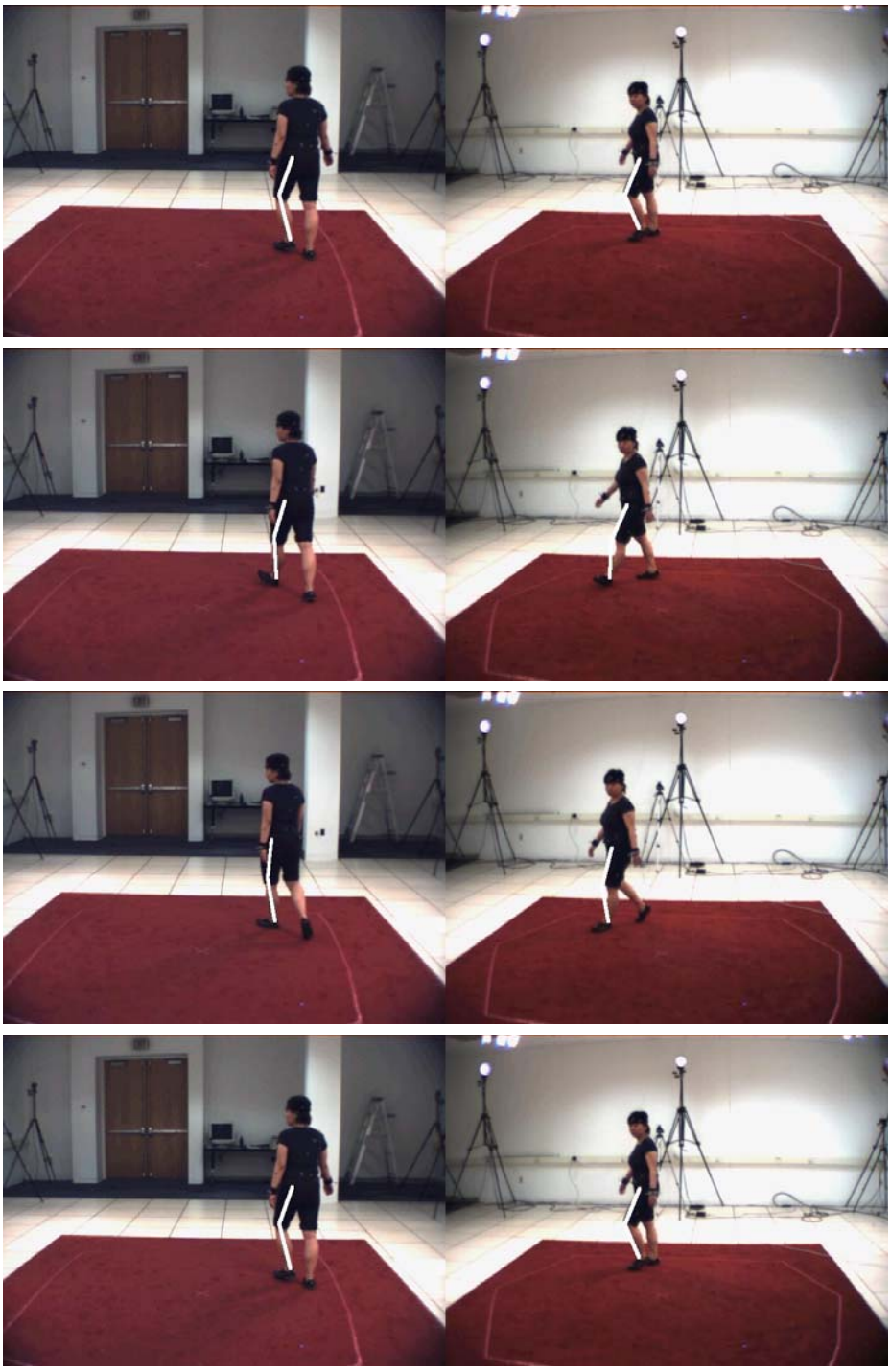

Figura 4.14: Seqüència *walking* de l'HumanEva per el subjecte 1 aplicant IK .

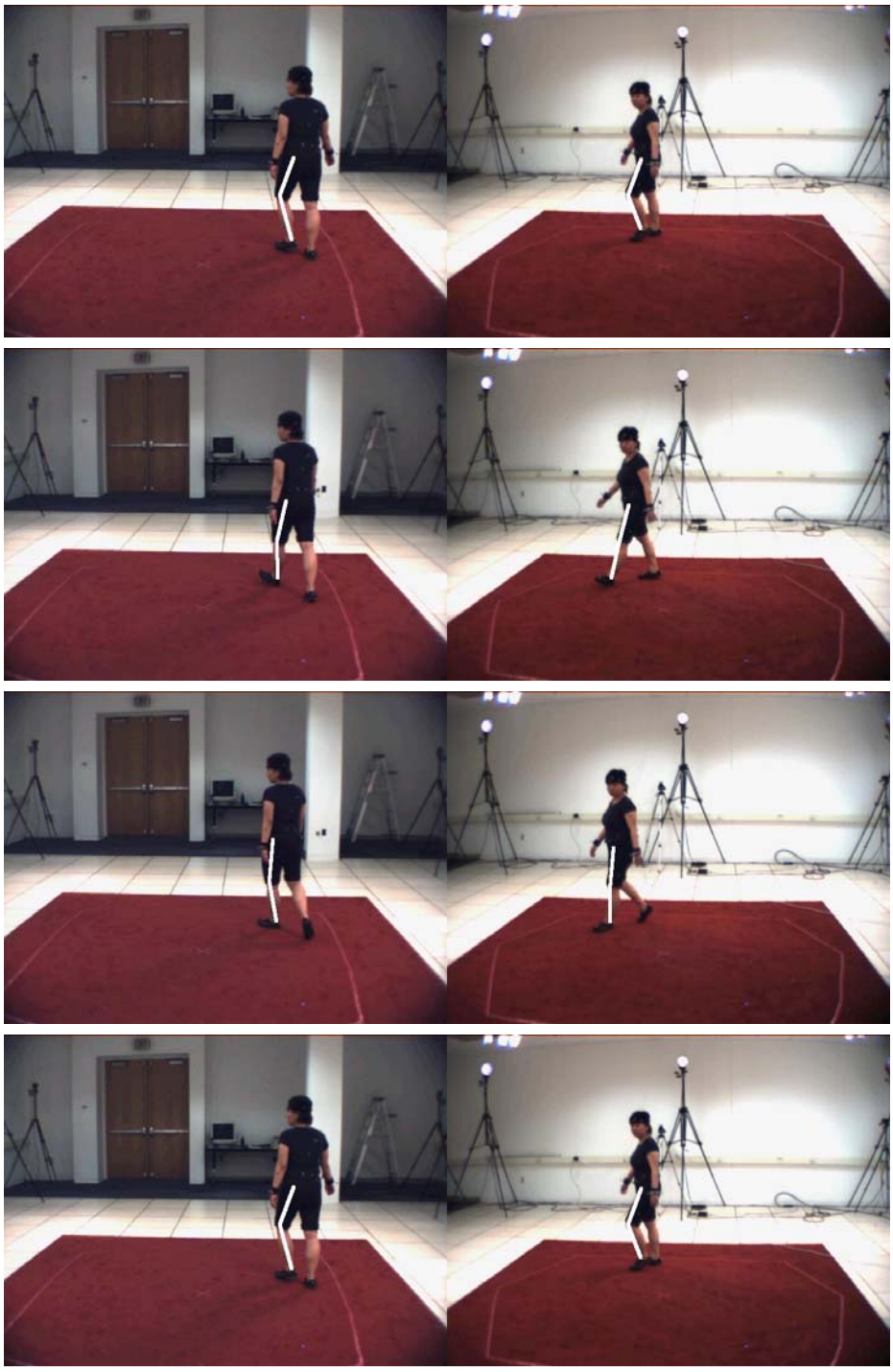

Figura 4.15: Seqüència *walking* de l'HumanEva per el subjecte 1 aplicant ibIK.

En el gràfic de la Figura 4.16, per la seqüència *walking*, es mostra l'error de l'estimació del genoll al aplicar cinemàtica inversa sense l'ajuda de la restricció basada en la imatge i es compara amb l'error de l'estimació de la posició del genoll al aplicar cinemàtica inversa amb una restricció basada en la imatge. Es pot observar que en la majoria dels casos l'estimació és millor amb ibIK.

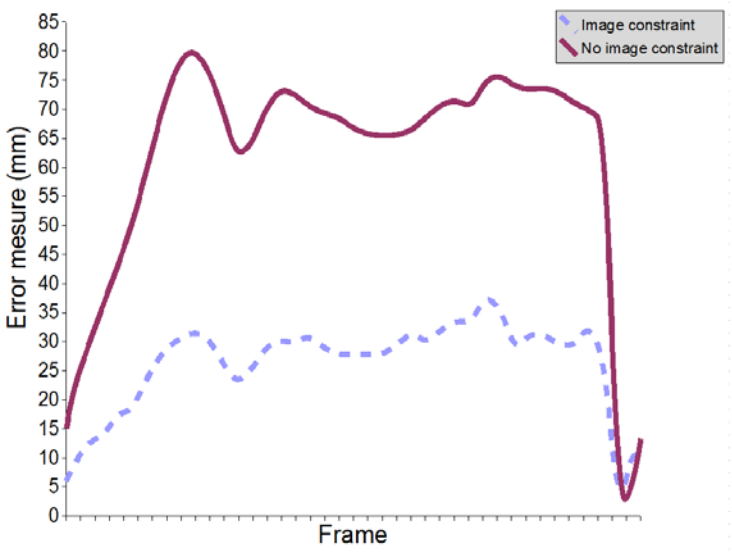

Figura 4.16: Error de l'estimació del genoll per cada imatge de la seqüència *walking*.

4.5 Resum

En aquest capítol s'ha presentat com afegir un restricció basada en la imatge en la formulació de la cinemàtica inversa, que consisteix en un criteri que intenta guiar la cadena cinemàtica a la projecció del cos dins la imatge. D'aquesta forma, s'eviten configuracions impossibles de la cadena.

Experiments amb imatges sintètiques han demostrat que aquesta aproximació

funciona correctament. L'avaluació del sistema amb imatges reals, ha permès demostrar que l'aproximació també funciona en entorns no controlats. L'error d'uns 2 centímetres, diferència mitja entre el valors reals i obtinguts per l'aproximació presentada, pot ser considerat suficientment petit per permetre el seu ús en aplicacions de captura del moviment.

Capítol 5

Conclusions

. . . i tot s'acaba.

Actualment s'està realitzant un esforç considerable en la investigació de mètodes pel reconeixement del moviment de les persones, a causa del potencial que representa la seva aplicació en el context de la interacció persona-ordinador. La majoria de treballs anteriors es basen en utilitzar directament els valors de la imatge per fer el reconeixement, degut a la dificultat de trobar les posicions 3D de les parts del cos en temps real. En aquesta memòria s'ha presentat un mètode que permet obtenir els moviments de l'usuari en un espai 3D. El principal avantatge del sistema presentat, és que evita específicament mètodes intrusius tal com són els marcadors, ja que està basat en visió. Aquest fet, permet a l'usuari realitzar un àmplia gamma de moviments. A més, el procés complet es du a terme en temps real per aconseguir una interacció fiable.

Mitjançant l'ús d'un model basat en la cinemàtica inversa, el sistema és potencialment més precís i robust en quan a oclusions que els enfocaments basats en la detecció de canvis ens els píxels. Això, és degut al model de restriccions addicionals que poden ser utilitzades per resoldre qualsevol discrepància entre les posicions mesurades i les posicions predites. Per avaluar el sistema, en primer lloc s'ha validat l'algorisme de visió que localitza la posició 3D de les mans, comparant els resultats obtinguts pel sistema de visió amb els retornats per un dispositiu de posicionament per ultrasons. L'error quadràtic mitjà entre el dispositiu i el sistema de visió ha estat d'uns 55 mm en el cas del moviments de prova més complexos i de 4.8 mm en el cas de posicions estàtiques. A continuació, s'ha provat el sistema complet amb experiments per mesurar l'exactitud de l'estimació de les articulacions internes, amb un experiment on els usuaris havien de fer diversos moviments predefinits. S'ha comparat la posició dels colzes de dues seqüències anotades (una simple i una complexa) amb les posicions estimades pel sistema de captura presentat. En el cas de la seqüència simple l'error quadràtic mitjà ha estat de 50 mm i en el cas de la seqüència complexa de 120 mm. La qualitat dels resultats és suficient per l'objectiu proposat, que consistia en obrir camí en l'explotació d'una forma de captura no invasiva i coherent amb l'espai de postures del cos complet per interaccions 3D en temps real.

A partir del mètode de captura del moviment presentat, en aquest treball s'ha presentat una proposta per reconèixer els gestos que realitza l'usuari en cada moment, amb l'objectiu de realitzar una interacció natural. La novetat recau en la representació de la postura, que permet a la interfície generalitzar sobre les diferències de la forma del cos en la població dels usuaris. Per a l'aprenentatge i el

reconeixement, s'ha utilitzat la tècnica no paramètrica del *k-veï més proper*. Els experiments han demostrat que, des d'un punt de vista pràctic, aquesta tècnica de classificació és apropiada per problemes del món real, a causa de la seva simplicitat en l'aprenentatge i la classificació en temps real. Endemés, el sistema s'adapta a la forma particular en què cada usuari realitza els moviments, donant llibertat a l'usuari i evitant que hagi d'aprendre prèviament els moviments que pot realitzar. Per avaluar la proposta de reconeixement, aquesta ha estat provada en una aplicació en temps real, el control basat en el moviment d'un videojoc per diferents usuaris que mai havien utilitzat l'aplicació. Els resultats han mostrat, que la proposta obté un rati de reconeixement superior al 95% per a moviments reconeguts que són els que envien esdeveniments al videojoc. L'anàlisi dels resultats mostraren que la majoria de moviments no reconeguts eren deguts a errors del sistema de Visió-PIK en l'estimació de les articulacions del cos de l'usuari. Al no disposar d'informació que permetés localitzar les articulacions internes, es forçava a la cinemàtica inversa a estimar arbitràriament la configuració d'aquestes articulacions.

Per això, en aquest treball també s'ha presentat com afegir un restricció basada en la imatge en la formulació de la cinemàtica inversa amb l'objectiu de resoldre aquest problema. S'ha proposat un criteri que intenta guiar la cadena articulada a la projecció del cos dins la imatge. D'aquesta forma, s'eviten configuracions impossibles de la cadena respecte l'usuari, com per exemple que el colze estigui fora de la projecció del cos de l'usuari. Experiments amb imatges sintètiques han demostrat que aquesta aproximació funciona correctament, i permet resoldre situacions difícils que succeeixen quan hi ha moviments on els *end-effectos* no hi estan implicats. Cal

afegir, que també s'ha avaluat el sistema utilitzant imatges reals, incloent seqüències d'una coneguda base de dades de moviment humà, per poder calcular resultats quantitatius. L'error calculat, d'uns 2 centímetres, pot ser considerat suficientment petit per permetre el seu ús en aplicacions de captura del moviment.

Sobre la implementació del sistema, que inclou tots els algorismes presentats, s'ha realitzat utilitzant Visual C++ amb les llibreries OpenCV [12] i s'ha testejat en un context d'interacció de temps real sobre un Intel Core2 QUAD Q6600 amb Windows Vista, obtenint un rati de 19 imatges per segon. Per tant,també és possible, l'ús en aplicacions d'interacció persona-ordinador.

Com a treball futur, es pretén generalitzar l'enfocament per incloure més tasques basades en la imatge utilitzant l'estratègia de prioritats. En aquest sentit, seria possible usar models més complexos del cos humà amb l'objectiu d'obtenir estimacions millors. Endemés, ja s'ha començat a transferir els resultats d'aquests treballs en àrees de rehabilitació i telerehabilitació [79], transferència que es vol continuar en el futur.

5.1 Publicacions i contribucions

Les propostes d'aquest treball, s'han publicades per estar a disposició de la comunitat científica. A més, els articles són considerats la forma d'avaluar la novetat i la fiabilitat del treball científic. Amb l'objectiu de ressaltar la novetat introduïda per aquest treball, els articles publicats es detallen a continuació

5.1.1 Articles

- Jaume-i-Capó, A.; Varona, J.; González-Hidalgo, M.; Perales, F.J.; *Adding image constraints to inverse kinematics for human motion capture*; Journal on Advances in Signal Processing 2009; ISSN: 1687-6172 (Accepted, In Press July 2009). Impact Factor 1.055. Ranking 115/229.

- Varona, J.; Jaume-i-Capó, A.; Gonzàlez, J.; Perales, F.J.; *Toward natural interaction through visual recognition of body gestures in real-time*; Interacting With Computers 21 (1 – 2): 3 – 10, 2009; ISSN: 0953-5438. Impact Factor 1.103. Ranking 7/17

- Jaume-i-Capó, A.; Varona, J.; Perales, F.J.;*Representation of human postures for vision-based gesture recognition in real-time*; Lecture Notes in Artificial Intelligence 5085: 102 – 108, 2008, ISSN: 0302-9743

- Jaume-i-Capó, A.; Varona, J. and Perales, F.J.; *Real-Time Recognition of Human Gestures for 3D Interaction*; Lecture Notes in Computer Science 5098: 419 - 430 ; 2008; ISSN: 0302-9743

- González Hidalgo, M.; Jaume Capó, A.; Mir, A.; Nicolau Bestard, G.;*Analytical Simulation of B-Spline Surfaces Defomation*; Lecture Notes in Computer Science 5098: 338-348, 2008, ISSN: 0302-9743

5.1.2 Proceedings

- Antoni Jaume-i-Capó, Javier Varona, Francisco J. Perales, *Representation of human postures for vision-based gesture recognition in real-time*, Proceedings of GW2007, 7th International Workshop on Gesture in Human-Computer Interaction and Simulation, Lisbon (Portugal), ISBN: 978-972-8862-05.

- Jaume-i-Capó, A.; Varona, J.; Perales, F.J., *Interactive applications driven by human gestures*, SIACG'2006 - IBERO AMERICAN SYMPOSIUM IN COMPUTER GRAPHICS, Santiago de Compostel·la (ESPANYA).

- A. Jaume-i-Capó, J. Varona, M. González-Hidalgo, Fco. J. Perales, R. Mas, *Automatic Human Body Modeling for Vision-Based Motion Capture*, Journal of WSCG: International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG'06), Plzen, Czech Republic, ISBN: 80-86943-05-4.

- R. Boulic, J. Varona, B. Herbelin, L. Unzueta, A. Suescun, A. Jaume-i-Capó, F. Perales, D. Thalmann, *Vision-Based Comparative Study of Analytic and Numeric Inverse Kinematic Techniques for Recovering Arm Movements*, First International Enactive Workshop,Pisa (ITÀLIA), 2005.

- Bartomeu Mir, Antoni Salas, Antoni Jaume-i-Capó, Marta Bez, Francisco Perales, *Low Cost Avatars Animation System from Real Images Compliant MPEG4*, International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2005, ISBN: 80-903100-8-7.

- José Miguel Salles Dias, Antoni Jaume-i-Capo, Joan Carreras, Ricardo Galli, Manuel Gamito,*A4D: Augmented Reality 4D System for Architecture and Building Construction*, CONVR2003 - Conference on Construction Applications of Virtual Reality, Blacksbug, Virginia (ESTATS UNITS D'AMÈRICA), 2003.

5.1.3 Projectes

Les tècniques i metodologies desenvolupades han estat usades al següent conjunt de projectes:

- *PROTOTIPOS DE INTERACCION NATURAL MEDIANTE INTERFACES ENACTIVAS BASADAS EN ENTRADAS VISUALES (PINes)*, Programa Nacional de Tecnologías de la Información y las Comunicaciones, TIN2007-67896: 2007 – 2010, IP: Javier Varona

- *Red Temática en Procesamiento de la Señal Audio-Visual en Interfaces Multimodales Avanzados*, Programa Nacional de Tecnologías de la Información y las Comunicaciones, TIN2006-26901E: 2006 – 2008, IP: Nicolás Pérez de la Blanca

- *Encuentros sobre e-inclusión y rehabilitación usando interfaces de usuario basadas en visión*, Programa Nacional de Cooperación Internacional en Ciencia y Tecnología, C/020306/08: 2009 – 2010, IP: Antoni Jaume-i-Capó

- *Acción complementaria sobre interfaces naturales basadas en visión para dispositivos móviles*, Programa Nacional de Cooperación Internacional en Ciencia y Tecnología, B/016677/08: 2009 – 2010, IP: Antoni Jaume-i-Capó

- *DISENO DE UN SISTEMA DE RECOSTRUCCION 3D MEDIANTE CAMARAS ESTEREOSCOPICAS Y LUZ ESTRUCTURADA*, Agencia Estatal de Cooperación Internacional (AECI), A/7155/06: 2007 – 2008, IP: María José Abasolo

- *Integración de escenarios virtuales con agentes inteligentes 3D (INEVAI3D)*, Ministerio de Educación y Cultura, DGICYT TIN2004-07926-E: 2004–2007, IP: Francisco José Perales López

- *An automatic human model animation environment for augmented reality interaction (HUMODAN)*, Unión Europea, IST-2001-32202: 2002 – 2005, IP: Francisco José Perales López

- *DEGAP*, Unión Europea, IPS-2001-42114: 2002 – 2004, IP: Miguel Dias

5.1.4 Estades en centres de recerca

- Computer Graphics and Multimedia Laboratory. ADETTI group in ISCTE. Lisboa. Portugal. From the 20^{th} December 2008, to the 31^{st} March 2009.

Capítol 6

Conclusions in English

Nowadays, there is a considerable effort in the research of human motion recognition methods due to its potential application in huma-computer interaction. The majority of the previous works are based on using directly image values for recognition due to the difficulty of finding 3D body poses in real-time. We have presented an approach to obtain user's motion in a 3D-space. The main advantage of our system is that we avoid specifically invasive methods such as markers and that we allow the user to perform a broad range of motions, because is vision-based. Moreover, the whole process is done in real-time to achieve a reliable interaction.

By using an inverse kinematics based model, the system is potentially more accurate and robust to occlusion effects than approaches based on detection of pixel changes. This is because the model provides additional constraints that can be used to resolve any discrepancies between measured and predicted positions. To evaluate the system, first we have tested the vision algorithm, that locates the 3D positions

of the hands. We have compared the results obtained by the vision system with a ultrasound positioning device. The mean-square error between the device and the vision system has been about 55 mm in the case of complex movements and about 4.8 mm in the case of static positions. The, we have tested the complete system with experiments to measure the accuracy of the internal joints estimations with an experiment where the users have to do several predefined motions. We have compared the position of the elbows of two annotated sequences (simple and complex) with the estimated positions computed by the presented capture system. In the case of the simple sequence mean-square error was 50 mm and for the complex sequence was 120 mm. The quality of the results is sufficient for our objective, which is to open the way to exploit a non-invasive wide and coherent full-body postural space for real-time 3D interactions.

From the motion capture method presented, in this memory we also have presented an approach to recognize the gestures that the user performances in each moment, with the goal of natural interaction. The novelty lies in the representation of pose that allows the interface to generalize over body shape differences in the population of users. Our approach is original and it could be extended to represent more complex gestures and human activities. The complete system has been tested in a real-time application, a motion-based videogame control. For learning and recognition we used the non-parametric technique of the k-nearest neighbor. Experiments have shown that, from a practical point of view, this technique of classification is appropriated for real world problems due to its simplicity in learning and on-line classification. Besides, the system adapts itself to each particular user's

way of performing motions, avoiding a previous user's off-line training to learn the motions that can be recognized by the system. The complete system has been tested in a real-time application, a motion-based videogame control by different users. The results have shown that the recognition ratio is greater than 95% for recognized motions. The analysis of the results showed that the most non-recognized motions are due to errors of the Vision-PIK estimation of the user's body joints, because no information is available to locate the internal joints and this forced the IK approach to make a somewhat arbitrary decision about what was the optimal angle for these joints.

For these reasons, we also have presented how to add image constraints to the inverse kinematics formulation in order to solve this problem. We have proposed a criterion that tries to guide the articulated chain to the body projection into the image. In this way, impossible chain configurations are avoided. Experiments using synthetic images show how this approximation performs correctly and, how to solve difficult situations that occur when there are motions that do not imply to the end-effectors. Besides, we have evaluated our approach using real images, including sequences of a known human motion database in order to compute quantitative results. The computed error, about 2 centimeters, can be considered as sufficiently small to permit its use in motion capture applications.

In addition, the complete system, including all the algorithms presented, has been implemented in Visual C++ using the OpenCV libraries [12] and it has been tested in a real-time interaction context on an Intel Core2 QUAD Q6600 under Windows Vista, obtaining a performance of 19 frames per second. Therefore, its use in human-

computer interaction applications is also possible.

As future work, we plan to generalize this approach to include more tasks by using the priority strategy. In this way, it would be possible to use more complex models of the human body in to order to achieve better estimations. Moreover, we have started to apply this work in the areas of rehabilitation and telerehabilitation [79]

6.1 Publications and contributions

As part of the scientific method and scientific process, the discovered improvements should be published to be available for the scientific community, which can correct or take advantage of the acquired experience. Then, papers are considered the way to evaluate the novelty and reliability of a scientific work. In order to remark the novelty introduced by this work, the papers are detailed next

6.1.1 Journals

- Jaume-i-Capó, A.; Varona, J.; González-Hidalgo, M.; Perales, F.J.; *Adding image constraints to inverse kinematics for human motion capture*; Journal on Advances in Signal Processing 2009; ISSN: 1687-6172 (Accepted, In Press July 2009). Impact Factor 1.055. Ranking 115/229.

- Varona, J.; Jaume-i-Capó, A.; Gonzàlez, J.; Perales, F.J.; *Toward natural interaction through visual recognition of body gestures in real-time*; Interacting With Computers 21 (1 – 2): 3 – 10, 2009; ISSN: 0953-5438. Impact Factor 1.103. Ranking 7/17

- Jaume-i-Capó, A.; Varona, J.; Perales, F.J.;*Representation of human postures for vision-based gesture recognition in real-time*; Lecture Notes in Artificial Intelligence 5085: 102 – 108, 2008, ISSN: 0302-9743

- Jaume-i-Capó, A.; Varona, J. and Perales, F.J.; *Real-Time Recognition of Human Gestures for 3D Interaction*; Lecture Notes in Computer Science 5098: 419 - 430 ; 2008; ISSN: 0302-9743

- González Hidalgo, M.; Jaume Capó, A.; Mir, A.; Nicolau Bestard, G.;*Analytical Simulation of B-Spline Surfaces Defomation*; Lecture Notes in Computer Science 5098: 338-348, 2008, ISSN: 0302-9743

6.1.2 Proceedings

- Antoni Jaume-i-Capó, Javier Varona, Francisco J. Perales, *Representation of human postures for vision-based gesture recognition in real-time*, Proceedings of GW2007, 7th International Workshop on Gesture in Human-Computer Interaction and Simulation, Lisbon (Portugal), ISBN: 978-972-8862-05.

- Jaume-i-Capó, A.; Varona, J.; Perales, F.J., *Interactive applications driven by human gestures*, SIACG'2006 - IBERO AMERICAN SYMPOSIUM IN COMPUTER GRAPHICS, Santiago de Compostel·la (ESPANYA).

- A. Jaume-i-Capó, J. Varona, M. González-Hidalgo, Fco. J. Perales, R. Mas, *Automatic Human Body Modeling for Vision-Based Motion Capture*, Journal of WSCG: International Conference in Central Europe on Computer Graphics,

Visualization and Computer Vision (WSCG'06), Plzen, Czech Republic, ISBN: 80-86943-05-4.

- R. Boulic, J. Varona, B. Herbelin, L. Unzueta, A. Suescun, A. Jaume-i-Capó, F. Perales, D. Thalmann, *Vision-Based Comparative Study of Analytic and Numeric Inverse Kinematic Techniques for Recovering Arm Movements*, First International Enactive Workshop,Pisa (ITÀLIA), 2005.

- Bartomeu Mir, Antoni Salas, Antoni Jaume-i-Capó, Marta Bez, Francisco Perales, *Low Cost Avatars Animation System from Real Images Compliant MPEG4*, International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2005, ISBN: 80-903100-8-7.

- José Miguel Salles Dias, Antoni Jaume-i-Capo, Joan Carreras, Ricardo Galli, Manuel Gamito,*A4D: Augmented Reality 4D System for Architecture and Building Construction*, CONVR2003 - Conference on Construction Applications of Virtual Reality, Blacksbug, Virginia (ESTATS UNITS D'AMÈRICA), 2003.

6.1.3 Projects

In addition, the techniques and methodologies developed have been used in a large set of projects:

- *PROTOTIPOS DE INTERACCION NATURAL MEDIANTE INTERFACES ENACTIVAS BASADAS EN ENTRADAS VISUALES (PINes)*, Programa Nacional de Tecnologías de la Información y las Comunicaciones, TIN2007-67896: 2007 – 2010, IP: Javier Varona

- *Red Temática en Procesamiento de la Señal Audio-Visual en Interfaces Multimodales Avanzados*, Programa Nacional de Tecnologías de la Información y las Comunicaciones, TIN2006-26901E: 2006 – 2008, IP: Nicolás Pérez de la Blanca

- *Encuentros sobre e-inclusión y rehabilitación usando interfaces de usuario basadas en visión*, Programa Nacional de Cooperación Internacional en Ciencia y Tecnología, C/020306/08: 2009 – 2010, IP: Antoni Jaume-i-Capó

- *Acción complementaria sobre interfaces naturales basadas en visión para dispositivos móviles*, Programa Nacional de Cooperación Internacional en Ciencia y Tecnología, B/016677/08: 2009 – 2010, IP: Antoni Jaume-i-Capó

- *DISENO DE UN SISTEMA DE RECOSTRUCCION 3D MEDIANTE CAMARAS ESTEREOSCOPICAS Y LUZ ESTRUCTURADA*, Agencia Estatal de Cooperación Internacional (AECI), A/7155/06: 2007 – 2008, IP: María José Abasolo

- *Integración de escenarios virtuales con agentes inteligentes 3D (INEVAI3D)*, Ministerio de Educación y Cultura, DGICYT TIN2004-07926-E: 2004–2007, IP: Francisco José Perales López

- *An automatic human model animation environment for augmented reality interaction (HUMODAN)*, Unión Europea, IST-2001-32202: 2002 – 2005, IP: Francisco José Perales López

- *DEGAP*, Unión Europea, IPS-2001-42114: 2002 – 2004, IP: Miguel Dias

6.1.4 Research stays abroad

- Computer Graphics and Multimedia Laboratory. ADETTI group in ISCTE. Lisboa. Portugal. From the 20^{th} December 2008, to the 31^{st} March 2009.

Apèndix A

Modelat automàtic del cos de l'usuari

En la captura del moviment basada en visió, un punt important és el model del cos de l'usuari. Aquest model ha de ser precís per representar els moviments mitjançant postures del cos, però també ha de ser simple perquè el problema sigui tractable i obtenir un feedback en temps real de l'aplicació. Habitualment, aquest model es construït prèviament i pot ser modelat a partir d'imatges de l'usuari [58]. Les tècniques més comunes per modelar són els visual-hulls [14, 17]. Tots aquests models tenen una aparença realista però no són suficientment exactes per el seu ús en aplicacions en temps real. En aquest treball esteim interessats en construir models menys exactes, però que representin suficientment bé els moviments de l'usuari. Aleshores, s'intentarà modelar l'estructura cinemàtica de l'usuari [30].

Endemés, les propostes basades en visió es basen en la coherència temporal dels

moviments de l'usuari. Aquest fet implica conèixer la postura prèvia de l'usuari i la seva postura inicial. Això és, el model del cos ha de ser inicialitzat en la primera captura (frame). Aquesta inicialització consisteix en trobar les posicions 3D de les articulacions de l'usuari en la primera imatge. Actualment, una pràctica comuna dels treballs basats en visió és solucionar aquest problema mitjançant l'anotació manual [13, 21].

Per això, proposam una inicialització automàtica per al nostre sistema de captura del moviment. L'algorisme que proposam es basa en analitzar la forma del cos de l'usuari projectada en les imatges, o sigui la seva silueta. La idea principal es tallar cada silueta en diferents parts del cos assumint que l'usuari està en una postura predefinida. Conseqüentment, a partir d'aquests talls podem estimar les posicions 3D de les articulacions de l'usuari. Aleshores, també podem construir el model cinemàtic humà de l'usuari.

Una vegada que la forma humana s'ha obtinguda, es necessari tallar la silueta en diferents parts per trobar les articulacions. D'acord amb la intuïció humana sobre les parts, la segmentació entre parts es produeix en la curvatura mínima negativa (NMC) de la silueta, seguint la regla mínima de Hoffman i Richards: *Per qualsevol silueta, totes les curvatures mínimes negatives de la seva forma que la limiten, són límits entre les parts* [31].

Aleshores, és necessari trobar els punts de la NMC de la forma. És possible calcular la curvatura de la forma directament mitjançant diferències finites damunt la forma discreta. Tanmateix, aquest càlcul local de la curvatura no és robust en imatges. Per aquest motiu, s'ha de parametritzar la forma del cos humà usant una

interpolació B-Spline.

Una B-Spline de grau p és una corba paramètrica, on cada component és una combinació lineal de funcions base de grau p. A partir de la imatge obtenim un conjunt de punts del contorn $Q_k, k = 0, ...n$ i interpolam aquests punts amb una B-Spline cúbica [52]. En la Figura A.1 és pot veure com la interpolació B-Spline suavitza la silueta humana.

Figura A.1: Interpolació B-Spline d'una silueta humana.

Utilitzant la parametrització de la forma amb B-Spline és possible calcular analíticament les derivades parcials fins a ordre dos per obtenir els valors de la curvatura al llarg de la forma. En la Figura A.2 es mostren els valors màxims i mínims de la curvatura.

No obstant, la regla mínima només obliga que els talls passin a través dels punts de la NMC, però no dirigeix la selecció de talls. Per altra banda en [62] observaren que quan els punts es poden unir en més d'una forma per descompondre la silueta,

Figura A.2: Mínims (en blanc)i maxims (en verd) de la curvatura.

la visió humana prefereix l'esquema de tall que usa el tall més curt. Això condueix a la Regla del tall més curt (short-cut rule) que requereix que un tall:

- Sigui una línia recta.
- Creui un eix de simetria local.
- Uneixi dos punts en el contorn de la silueta, tals que com a mínim un d'ells tengui curvatura negativa.
- Sigui el més curt si hi ha varis talls que competeixen.

Per altra banda, si es coneix la postura de l'usuari es possible predir on són els tall. Amb l'objectiu d'obtenir més fàcilment els talls principals, es requereix que l'usuari estigui en una postura predefinida adequada per trobar totes les articulacions del nostre model del cos, veure Figura A.3.

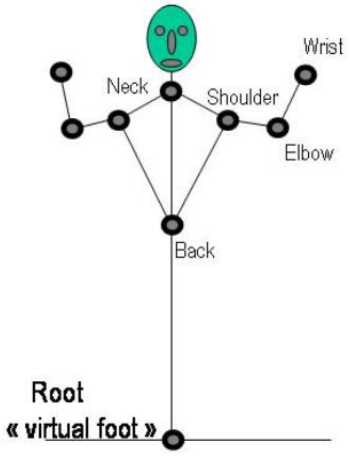

Figura A.3: Postura inicial.

Estudiant la postura inicial de la Figura A.3, s'ha detectat que els punts mínims i màxims de la curvatura mínima estan devora les articulacions que volem trobar. Aquest fet és clarament visible en la Figura A.2. Aleshores, d'acord amb la short-cut rule i tenguent en compte la postura inicial de l'usuari es proposen les següents normes per descompondre la forma humana:

- L'esquena (Back) està al punt NMC amb la component y més petita.
- El coll (Neck) està situat en el punt mig del tall que formen el dos punts amb la NMC amb la component y més grossa.
- Contruir l'eix principal del cos unint l'esquena i el coll. Aquest divideix la forma del cos en dues parts.
- Les espatlles (Shoulders) es col·loquen en el punt mig de tal que forma el punt

del NMC amb la component y més gros de la part dreta/esquerra, i el punt de la NMC amb la component y més petita de la part dreta/esquerra, excepte el punt de l'esquena.

- Els colzes (Elbows) està situat en el punt mig del tall entre el punt de la NMC amb la component x més petita/gran i el punt de la PMC amb la component x més gran/petita

Figura A.4: Talls del cos.

On x i y es refereix a les coordenades horitzontals i verticals de la imatge respectivament. Aplicant aquestes normes s'obté la descomposició de la silueta de l'usuari com es mostra a la Figura A.4. La Figura A.5 mostra el model humà a partir dels talls.

Figura A.5: Model del cos generat.

Conseqüentment, s'estimen les posicions de les espatlles i els colzes com el punt mig del tall. Una vegada que s'han calculat les posicions 2D en cada imatge del sistema estèreo, es pot estimar la seva posició 3D usant el mètode de la triangulació del punt mig. Amb aquest mètode, la posició 3D és calculada projectant cada posició 2D de l'articulació sobre cada imatge a l'infinit i calculant les seves coordenades 3D com el punt més proper d'aquestes dues línees [67].

Finalment, els canells s'han d'estimar per completar el model del cos. No obstant, no es poden detectar usant el mètode d'anàlisis de la forma que hem proposat. Per detectar les posicions 3D dels canells, s'utilitzen les el·lipses 2D calculades en el procés del *Seguiment 2D*, la posició 3D de les mans calculades en el procés de *Seguiment 3D* i la posició 3D prèvia dels colzes. Amb aquestes dades, es cerca en la imatge la intersecció entre una línia 2D, definida per la retroprojecció que correspon en la

unió del posició del colze i la posició de la mà, i la el·lipse 2D. Finalment, a partir de les posicions 2D d'aquestes interseccions, els canells, es calculen les coordenades 3D triangulant [71]. Alguns resultats de procés de localització dels canells és mostren a la Figura A.6.

Figura A.6: Estimació de la posició dels canells.

Bibliografia

[1] Norman I. Badler, Michael J. Hollick, and John P. Granieri. Real-time control of a virtual human using minimal sensors. *Presence*, 2(1):82 – 86, 1993.

[2] Paolo Baerlocher and Ronan Boulic. An inverse kinematics architecture enforcing an arbitrary number of strict priority levels. *The Visual Computer*, 20(6):402–417, 2004.

[3] Donald. G Bailey. An efficient euclidean distance transform. In *Combinatorial Image Analysis, IWCIA 2004*, pages 394–408, 2004.

[4] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, Mar 2001.

[5] R. Bodenheimer, C. Rose, S. Rosenthal, and J.Pella J. The process of motion capture: dealing with the data. In Eurographics Association, editor, *Computer animation and simulation*, pages 3 – 18, 1997.

[6] Richard A. Bolt. 'put-that-there': Voice and gesture at the graphics interface. In *SIGGRAPH '80: Proceedings of the 7th annual conference on Computer*

graphics and interactive techniques, pages 262–270, New York, NY, USA, 1980. ACM Press.

[7] R. Boulic, P. Baerlocher, I. Rodríguez, M. Peinado, and D. Meziat. Virtual worker reachable space evaluation with prioritized inverse kinematics. In *35th International Symposium on Robotics (on-line)*, 2004.

[8] R. Boulic, J. Varona, B. Herbelin, A. Suescun L. Unzueta, F. Perales A. Jaume-i Capó, and D. Thalmann. Vision-based comparative study of analytic and numeric inverse kinematic techniques for recovering arm movements. In *On-line Abstract Proceedings of the First International Enactive Workshop*, Pisa, 2005.

[9] R. Boulic, J. Varona, L. Unzueta, M. Peinado, A. Suescun, and F. Perales. Evaluation of on-line analytic and numeric inverse kinematics approaches driven by partial vision input. *Virtual Reality (online)*, 10(1):48–61, 2006.

[10] Ronan Boulic, Manuel Peinado, and Benoît Le Callennec. Challenges in exploiting prioritized inverse kinematics for motion capture and postural control. In *Gesture Workshop*, pages 176–187, 2005.

[11] Gary R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, 1(Q2):1 – 15, 1998.

[12] G.R. Bradski and V. Pisarevsky. Intel's computer vision library: applications in calibration, stereo segmentation, tracking, gesture, face and object recognition. In *Proceedings IEEE IEEE Conference on Computer Vision and Pattern*, volume 2, pages 796–797, 2000.

[13] Christoph Bregler, Jitendra Malik, and Katherine Pullen. Twist based acquisition and tracking of animal and human kinematics. *Int. J. Comput. Vision*, 56(3):179–194, 2004.

[14] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3):569–577, 2003.

[15] Jinxiang Chai and Jessica K. Hodgins. Performance animation from low-dimensional control signals. *ACM Transactions on Graphics*, 24:686 – 696, 2005.

[16] F. Chaumette and S. Hutchinson. Visual servo control, Part II: Advanced approaches. *IEEE Robotics and Automation Magazine*, 14:109–118, 2007.

[17] Kong Man Cheung, Simon Baker, and Takeo Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 77 – 84, June 2003.

[18] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[19] Dorin Comaniciu and Visvanathan Ramesh. Robust detection and tracking of human faces with an active camera. In *Proceedings of the Third IEEE Interna-*

tional Workshop on Visual Surveillance (VS'2000), pages 11–18, Washington, DC, USA, 2000. IEEE Computer Society.

[20] Carl D. Crane, III and Joseph Duffy. *Kinematic Analysis of Robot Manipulators*. Cambridge University Press, New York, NY, USA, 1998.

[21] Jonathan Deutscher and Ian Reid. Articulated body motion capture by stochastic search. *Int. J. Comput. Vision*, 61(2):185–205, 2005.

[22] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[23] A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proceedings. Ninth IEEE International Conference on Computer Vision, 2003.*, volume 2, pages 726–733, Oct. 2003.

[24] A. Elgammal, V. Shet, Y. Yacoob, and L.S. Davis. Learning dynamics for exemplar-based gesture recognition. In *Proceedings of Computer Vision and Pattern Recognition (CVPR'03)*, pages 571–578, 2003.

[25] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2003.

[26] W.T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma. Computer vision for computer games. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 100–105, Oct 1996.

[27] K Grochow, SL Martin, A Hertzmann, and Z Popovic. Style-based inverse kinematics. *ACM Transactions on Graphics*, 23(3):522 531, 2004.

[28] R. Hartley and P. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157, 1997.

[29] J. Heikkila and O. Silven. A four-step camera calibration procedure with implicit image correction. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1106–1112, Jun 1997.

[30] A. Hilton, M. Kalkavouras, and G. Collins. 3d studio production of animated actor models. *Vision, Image and Signal Processing, IEEE Proceedings* -, 152(4):481–490, Aug. 2005.

[31] D. D. Hoffman and W. A. Richards. Parts of recognition. *Readings in computer vision: issues, problems, principles, and paradigms*, pages 227–242, 1987.

[32] T. Horprasert, D. Harwood, and L. S. Davis. A robust background subtraction and shadow detection. In *Proceedings of the Asian Conference on Computer Vision*, 2000.

[33] Johanna Höysniemi, Perttu Hämäläinen, Laura Turkki, and Teppo Rouvi. Children's intuitive gestures in vision-based action games. *Commun. ACM*, 48(1):44–50, 2005.

[34] C.A. Klein, C. Chu-Jenq, and S. Ahmed. A new formulation of the extended jacobian method and its use in mapping algorithmic singularities for kinematically redundant manipulators. *Robotics and Automation, IEEE Transactions on*, 11(1):50–55, Feb 1995.

[35] Mark L. Knapp and Judith A. Hall. *Nonverbal Communication in Human Interaction*. Wadsworth Publishing, 6 edition, March 2005.

[36] A. Kojima, M. Izumi, T. Tamura, and K. Fukunaga. Generating natural language description of human behavior from video images. In *Proceedings. 15th International Conference on Pattern Recognition*, volume 4, pages 728–731 vol.4, 2000.

[37] J. Kovac, P. Peer, and F. Solina. Human skin color clustering for face detection. In *EUROCON 2003. Computer as a Tool. The IEEE Region 8*, volume 2, pages 144–148 vol.2, Sept. 2003.

[38] A. Liégeois. Automatic supervisory control of the configuration and behavior of multibody mechanisms. *Systems, Man and Cybernetics, IEEE Transactions on*, 7(12):868–871, Dec. 1977.

[39] A.A. Maciejewski. Dealing with the ill-conditioned equations of motion for articulated figures. *Computer Graphics and Applications, IEEE*, 10(3):63–71, May 1990.

[40] É. Marchand and N. Courty. Controlling a camera in a virtual environment. *The Visual Computer*, 18(1):1 – 19, 2002.

[41] Lars Mündermann, Stefano Corazza, and Thomas P Andriacchi. The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. *Journal of NeuroEngineering and Rehabilitation (on-line)*, 3(6), 2006.

[42] Thomas B. Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231 – 268, 2001.

[43] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90 – 126, 2006. Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour.

[44] Thomas B. Moeslund, Lars Reng, and Erik Granum. Finding motion primitives in human body gestures. In *Gesture in Human-Computer Interaction and Simulation: 6th International Gesture Workshop*, pages 133 – 144, 2006.

[45] T. Molet, R. Boulic, S. Rezzonico, and D. Thalmann. An architecture for immersive evaluation of complex human tasks. *IEEE Transactions on Robotics and Automation*, 15(3):475–485, Jun 1999.

[46] Yoshihiko Nakamura and Hideo Hanafusa. Inverse kinematic solutions with singularity robustness for robot manipulator control. *J. Dyn. Syst. Meas. Control*, 108:163–171, 1986.

[47] James O'Brien, Bobby Bodenheimer, Gabriel J. Brostow, Jessica K. Hodgins, and Jessica K. Hodgins. Automatic joint parameter estimation from magnetic motion capture data. In *Graphics Interface*, pages 53–60, 2000.

[48] R. G. O'Hagan, A. Zelinsky, and S. Rougeaux. Visual gesture interfaces for virtual environments. *Interacting with Computers*, 14(3):231 – 250, 2002.

[49] S.C.W. Ong and S. Ranganath. Automatic sign language analysis: a survey and the future beyond lexical meaning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(6):873–891, Jun 2005.

[50] Vladimir I. Pavlovic, Rajeev Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):677–695, 1997.

[51] J. Perales, J. Varona, J.M. Buades, and A. Jaume i Capó. Humodan: Deliverable 2.3. Technical report, IST Project 2001 32202, 2004.

[52] Les Piegl and Wayne Tiller. *The NURBS book*. Springer, 1997.

[53] R. Polana and R. Nelson. Detecting activities. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2–7, Jun 1993.

[54] Ramprasad B. Polana and Randal C. Nelson. Nonparametric recognition of nonrigid motion. Technical report, University of Rochester, Rochester, NY, USA, 1995.

[55] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes in C (2nd ed.): the art of scientific computing*. Cambridge University Press, 1992.

[56] Cen Rao, Alper Yilmaz, and Mubarak Shah. View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203–226, 2002.

[57] J.M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proceedings. Fifth International Conference on Computer Vision*, pages 612–617, Jun 1995.

[58] Fabio Remondino. 3-d reconstruction of static human body shape from image sequence. *Computer Vision and Image Understanding*, 93(1):65 – 85, 2004.

[59] Haibing Ren, Guangyou Xu, and SeokCheol Kee. Subject-independent natural action recognition. In *Proceedings. Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004.*, pages 523–528, May 2004.

[60] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[61] L. Sigal and M.J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, 2006.

[62] M. Singh, G. Seyranian, and D. Hoffman. Parsing silhouettes: The short-cut rule. *Perception and Psy.*, 61(4):636–660, 1999.

[63] Alvy Ray Smith and James F. Blinn. Blue screen matting. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 259–268, New York, NY, USA, 1996. ACM Press.

[64] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(12):1371–1375, Dec 1998.

[65] D.J. Sturman. Computer puppetry. *IEEE Computer Graphics and Applications*, 18(1):38–45, Jan/Feb 1998.

[66] Deepak Tolani, Ambarish Goswami, and Norman I. Badler. Real-time inverse kinematics techniques for anthropomorphic limbs. *Graphical Models*, 62(5):353 – 388, 2000.

[67] E. Trucco and A. Verri. *Introductory Techniques for 3D Computer Vision*. Prentic-Hall, 1998.

[68] M. Turk and M. Kolsch. *Emerging Topics in Computer Vision*, chapter Perceptual interfaces. Prentice Hall PTR, 2004.

[69] Raquel Urtasun, David J. Fleet, and Pascal Fua. 3d people tracking with gaussian process dynamical models. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 238–245, Washington, DC, USA, 2006. IEEE Computer Society.

[70] F. Varela, E. Thompson, E. Rosch, and Anand Rangarajan. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, 1993.

[71] Javier Varona, Jose Maria Buades Rubio, and Francisco J. Perales López. Hands and face tracking for vr applications. *Computers & Graphics*, 29(2):179–187, 2005.

[72] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, 2003.

[73] Greg Welch and Gary Bishop. An introduction to the kalman filter. Tr 95-041, University of North Carolina, 2006.

[74] C.R. Wren and A.P. Pentland. Understanding purposeful human motion. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, pages 19 – 25, 2000.

[75] Ying Wu and Thomas S. Huang. Vision-based gesture recognition: A review. In *GW '99: Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, pages 103–115, London, UK, 1999. Springer-Verlag.

[76] Ying Wu and T.S Huang. Human hand modeling, analysis and animation in the context of human computer interaction. In *IEEE Signal Processing Magazine, Special issue on Immersive Interactive Technology*, pages 6–10, 1999.

[77] Z. Zhang. A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(11):1330–1334, Nov 2000.

[78] Jianmin Zhao and Norman I. Badler. Inverse kinematics positioning using nonlinear programming for highly articulated figures. *ACM Trans. Graph.*, 13(4):313–336, 1994.

[79] Huiyu Zhou and Huosheng Hu. Human motion tracking for rehabilitation–a survey. *Biomedical Signal Processing and Control*, 3(1):1 – 18, 2008.

[80] Beiji Zou, Shu Chen, Cao Shi, and Umugwaneza Marie Providence. Automatic reconstruction of 3d human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking. *Pattern Recognition*, 42(7):1559 – 1571, 2009.