
Índice general

Índice general	1
1. Hardware y Software	2
1.1. Introducción	2
1.2. Software de procesamiento de imágenes	2
1.2.1. Lenguaje C	2
1.2.2. Librerías y recursos	2
1.2.2.1. OpenCV	2
1.2.2.2. IPOL	2
1.2.2.3. ITK e ImageMagik	2
1.3. Dispositivos móviles Apple	2
1.3.1. iPhone, iPad, iPod Touch	2
1.3.1.1. Comparación de plataformas.	2
1.3.1.2. Algunas características a detallar.	3
1.4. Software de Apple Inc.	5
1.4.1. Sistemas Operativos	5
1.4.2. Objective-C	5
1.4.3. Xcode: Herramientas y Librerías	6
1.4.3.1. Cocoa Touch Layer	7
1.4.3.2. Media Layer	8
1.4.3.3. Core Services	9
1.4.3.4. Core OS	10
1.4.3.5. Simulador	10
1.4.3.6. Instruments	11
1.5. Herramientas	12
1.5.1. GIT	12
1.5.2. GoogleCode	12
1.5.3. Github	12
2. Identificación	13
2.1. Introducción	13
2.2. Técnicas de identificación	13
2.3. QR	13
2.3.1. Identificadores QR: una realidad cotidiana.	13
2.3.2. ¿Qué son los QR?	13
2.3.3. Codificación y decodificación de códigos QR.	14
2.3.4. Expresiones artísticas con QRs.	15
2.4. SIFT	15

2.4.1. Detección de extremos en el espacio-escala.	16
2.4.2. Localización exacta de puntos clave.	16
2.4.3. Asignación de orientación.	18
2.4.4. Descriptor de puntos clave.	18
3. Detección	20
3.1. Tipos de características	20
3.2. Bordes y esquinas	20
3.2.1. Detector de bordes de Canny	20
3.2.2. Detector de bordes y esquinas de Harris	20
3.2.3. SUSAN Y FAST	20
3.3. Líneas y segmentos de línea	20
3.3.1. Detector de líneas de Hough	20
3.3.2. Detector de segmentos de línea: LSD	20
3.3.3. Detector de segmentos de línea: EDLines	20
3.4. Regiones y puntos de interés	20
3.4.1. FAST	20
3.4.2. Blobs	20
3.5. Descriptores	20
4. Marcadores	21
4.1. Sistemas basados en marcadores planos	21
4.1.1. ARToolKit	22
4.1.2. ARTag	23
4.2. Marcador QR	23
4.2.1. Estructura del marcador	24
4.2.2. Diseño	24
4.2.3. Parámetros de diseño	25
4.2.4. Diseños utilizados	27
4.3. Detección	27
4.3.1. Detección de segmentos de línea	27
4.3.2. Filtrado y agrupamiento de segmentos	28
4.3.3. Determinación de correspondencias	30
4.3.4. Detección robusta	33
4.3.5. Resultados	34
5. LSD: “Line Segment Detection”	35
5.1. Introducción	35
5.2. <i>Line-support regions</i>	36
5.3. Aproximación de las regiones por rectángulos	37
5.4. Validación de segmentos	37
5.5. Refinamiento de los candidatos	38
5.6. Optimización del algoritmo para tiempo real	39
5.6.1. Filtro Gaussiano	39
5.6.2. <i>Level-line angles</i>	41
5.6.3. Refinamiento y mejora de los candidatos	41
5.6.4. Algoritmo en precisión simple	42
5.6.5. Resultados	42
5.6.5.1. Filtro Gaussiano	42

5.6.5.2. <i>Line Segment Detection</i>	42
5.7. Conslusión	44
6. Modelo de cámara y estimación de pose monocular	46
6.1. Introducción	46
6.2. Modelo de cámara <i>pin-hole</i> [1]	47
6.2.1. Fundamentos y definiciones	47
6.2.2. Matriz de proyección	48
6.3. Distorsión introducida por las lentes	51
6.4. Métodos para la calibración de cámara	51
6.5. Calibración de cámaras utilizadas	54
6.5.1. Calibración <i>iPod Touch 4^{ta} generación</i>	54
6.5.2. Calibración <i>iPhone 4</i>	54
6.5.3. Calibración <i>iPad 2</i>	54
6.6. Problema de estimación de pose	55
6.6.1. <i>DLT(Direct Linear Transform)</i>	55
6.6.2. <i>PnP (Perspective-n-Point)</i>	56
6.6.3. <i>RANSAC(RANdom SAmple Consensus)</i>	56
6.6.4. <i>POSIT</i>	57
6.7. Representación de la pose de la cámara	58
6.7.1. Representación matricial	58
6.7.2. Ángulos de Euler	58
6.7.2.1. Orden de rotaciones	58
6.7.2.2. Cálculo de los ángulos de Euler	59
6.7.2.3. Gimbal lock	60
6.7.3. Cuaternios	61
7. POSIT: <i>POS</i> with <i>ITerations</i>	62
7.1. Introducción	62
7.2. POSIT clásico	62
7.2.1. Notación	62
7.2.2. SOP: <i>Scaled Ortographic Projection</i>	64
7.2.3. Ecuaciones para calcular la proyección perspectiva	64
7.2.4. Algoritmo	65
7.2.5. POSIT para puntos coplanares	66
7.3. SoftPOSIT	69
7.3.1. Modern POSIT	70
7.3.2. Cálculo de pose sin correspondencias	72
7.3.3. Matriz de asignación	72
7.3.4. Implementación	73
7.4. POSIT moderno para puntos coplanares	73
7.5. Resultados	74
8. Filtrado de Kalman para estimación de pose	78
8.1. Introducción	78
8.2. Filtro de Kalman	78
8.3. Kalman para suavizado	79
8.4. Kalman con sensores	79
8.5. Resultados	80

9. Rendering en iOS: ISGL3D	81
9.1. Introducción	81
9.2. Conceptos básicos de ISGL3D	82
9.3. FOV y ejes de ISGL3D	83
9.4. Primitivas de ISGL3D	83
9.5. Importación de modelos a ISGL3D.	86
9.6. Luz en ISGL3D	89
9.7. Método - <i>(void) tick:(float)dt</i>	89
9.8. ISGL3D en la aplicación	89
9.9. Conclusión	90
10. Casos de Uso	91
10.1. Introducción	91
10.2. Caso de uso “interactividad”	91
10.2.1. Comentarios sobre el caso de uso	91
10.2.2. Detalles constructivos	93
10.2.2.1. Objetos ISGL3D interactivos	93
10.2.2.2. Reproducción de audio en Objective-C	93
10.2.2.3. Dibujar en ISGL3D	94
10.3. Caso de uso “video”	94
10.3.1. Comentarios sobre el caso de uso	94
10.3.2. Detalles constructivos	95
10.3.3. <i>CGAffineTransform</i> y <i>CATransform3D</i>	95
10.3.4. Resolución de Homografía	97
10.4. Caso de uso “modelos”	98
10.4.1. Comentarios sobre el caso de uso	98
10.4.2. Detalles constructivos	98
11. Implementación	99
11.1. Introducción	99
11.2. Diagrama global de la aplicación	99
11.2.1. NavigationViewController	101
11.2.2. InicioViewController	102
11.2.3. UITableViewControllers	103
11.2.3.1. AutorTableViewController	103
11.2.3.2. CuadroTableViewController	104
11.2.3.3. CuadroTableViewCell	104
11.2.4. ReaderSampleViewController	105
11.2.5. ImagenServerViewController	105
11.2.6. ObraCompletaViewController	106
11.2.7. VistaViewController	106
11.2.8. DrawSign	107
11.2.9. TouchVista	109
11.2.10. Realidad Aumentada en ISGL3D	109
11.3. QR	111
11.3.1. Identificadores QR. Una realidad	111
11.3.2. ¿Qué son realmente los QR?	111
11.3.3. Codificación y decodificación de códigos QR	112
11.3.4. El QR en la aplicación	113

11.3.5. Buen gusto para los QR	113
11.4. Servidor	114
11.4.1. Creando el servidor	114
11.4.1.1. Servidor LAMP	114
11.4.1.2. Servidor en Mac OS X	114
11.4.1.3. Aspectos a mejorar del Servidor	115
11.5. SIFT	115
11.6. Comentarios finales sobre la implementación	116
Bibliografía	118

CAPÍTULO 1

Hardware y Software

1.1. Introducción

Introducción

1.2. Software de procesamiento de imágenes

Software de procesamiento de imágenes

1.2.1. Lenguaje C

Ventajas del Lenguaje C para procesamiento de imágenes

1.2.2. Librerías y recursos

1.2.2.1. OpenCV

1.2.2.2. IPOL

1.2.2.3. ITK e ImageMagik

1.3. Dispositivos móviles Apple

Al trabajar con Apple se cuenta con la ventaja de contar con pocas variantes en cuanto al Hardware utilizado. Básicamente existen tres familias de dispositivos en los que se puede desarrollar: iPhone, iPad y iPod Touch. Para cada variante de plataforma existen distintos modelos que hacen que algunas características importantes como la capacidad de procesamiento, la resolución de cámara o el tamaño de la pantalla entre otras puedan verse afectadas. A continuación se presenta brevemente cómo fue el surgimiento de cada uno de los dispositivos al mercado y se describen resumidamente las principales características.

1.3.1. iPhone, iPad, iPod Touch

1.3.1.1. Comparación de plataformas.

Sin dudas el iPhone fue uno de los saltos más grandes en el mundo tecnológico en los últimos años. Logró llenar el hueco que los PDAs de la década de los 90 no habían sabido completar y

comenzó a desplazar al invento que revolucionó el mercado de los contenidos de música, el iPod. Gracias a su pantalla táctil capacitiva de alta sensibilidad logró reunir todas las funcionalidades agregando solamente un gran botón y algunos extra para controlar volumen o desbloquear el dispositivo.

La primera generación del iPhone fue lanzada por Apple en Junio de 2007 en Estados Unidos, luego de una gran inversión de la operadora AT&T que exigía exclusividad de venta dentro de dicho país durante los siguientes cuatro años. La misma soportaba tecnología GSM cuatribanda y se lanzó en dos variantes de 4GB y 8GB de ROM. El segundo modelo lanzó como novedad el soporte de tecnología 3G cuatribanda y GPS asistido. Luego le siguieron el iPhone 3GS, 4, 4S y el 5, siendo este último, la sexta y última generación disponible al momento de la redacción de este trabajo.

Por su parte tanto el iPad como el iPod Touch también representaron un gran salto en el mundo de las plataformas y *Tablets*, agrandando las posibilidades de desarrollo y procesamiento. Como se dijo, de cada una de estas tres familias de dispositivos existen distintas versiones y modelos. Por eso, a continuación se muestra una tabla comparativa de determinadas características que son de interés a los efectos del presente proyecto.

Tabla 1.1: Comparativa de algunas plataformas Apple

	iPhone 4	iPhone 4s	iPod Touch 4G	iPad 2
ROM	8, 16 o 32 GB	16, 32 o 64 GB	8, 32 o 64 GB	16, 32 o 64 GB
RAM	512 MB	512 MB	256 MB	512 MB
SoC	Apple A4	Apple A5	Apple A4	Apple A5
CPU	1 GHz, ARM Cortex-A8	1 GHz, dual-core ARM Cortex-A9	800 MHz, ARM Cortex-A8	1 GHz dual-core ARM Cortex-A9
GPU	PowerVR SGX535 GPU	PowerVR SGX543MP2 (2-core) GPU	PowerVR SGX535 GPU	PowerVR SGX543MP2 (2-core) GPU
CÁMARA	Foto: 5.0 MP Video: 720p HD (30 fps)	Foto: 8.0 MP Video: 1080p HD (30 fps)	Foto: 0.7 MP Video: 720p HD (30 fps)	Foto: 0.7 MP Video: 720p HD (30 fps)
PANTALLA	Diagonal: 3.5" Pixels: 960x640 Densidad de Pixels: 326 ppi Multitáctil	Diagonal: 3.5" Pixels: 960x640 Densidad de Pixels: 326 ppi Multitáctil	Diagonal: 3.5" Pixels: 960x640 Densidad de Pixels: 326 ppi Multitáctil	Diagonal: 9.7" Pixels: 1024x768 Densidad de Pixels: 123 ppi Multitáctil
SENSORES	Girsóscopo de 3 ejes Acelerómetro Sensor de luz ambiente Sensor de proximidad	Girsóscopo de 3 ejes Acelerómetro Sensor de luz ambiente Sensor de proximidad	Girsóscopo de 3 ejes Acelerómetro Sensor de luz ambiente	Girsóscopo de 3 ejes Acelerómetro Sensor de luz

1.3.1.2. Algunas características a detallar.

Hay algunos comentarios respecto de la Tabla 1.1 que es bueno destacar. Primeramente, es importante decir que se eligieron esos cuatro dispositivos pues pareció de interés conocer al menos una plataforma de cada familia y dentro de las mismas se eligieron las que fueron utilizadas para desarrollar.

Uno de los puntos a evaluar es el **SoC**, que refiere a *System on Chip* por sus siglas en inglés. *System on Chip* es un concepto de los sistemas embebidos que refiere a la integración de todo lo necesario para poder correr un sistema operativo, en un solo circuito integrado. En contraposición a un microcontrolador que es capaz de realizar procesamiento más básico y menos potente, con poca interacción de usuario y menor flexibilidad, un *SoC* refiere a la idea de tener todo lo necesario para

desarrollar sobre la plataforma y poder hacer procesamiento sin tantas limitaciones. Básicamente cumplen funciones similares pero el *SoC* forma parte de una evolución de los microcontroladores, siendo de una complejidad mayor e integrado en un tamaño muy reducido buscando poco consumo y eficiencia de costos. Así entonces un *SoC* puede estar conformado por un microcontrolador y hardware adicional como procesadores de señal y bloques de memoria.

En la Figura 1.1 se ilustran los dos tipos de *SoC* de los dispositivos de la Tabla 1.1: Apple A4 y Apple A5. Algunos dispositivos que no figuran en la tabla como el *iPhone 5* o el *iPad 4* usan *SoCs* más recientes como el Apple A6 o Apple A6x respectivamente. La familia de *SoCs* Apple Ax, es la que la mencionada firma utiliza en todas sus plataformas, inclusive en el *Apple TV* y es manufacturada por Samsung. Estos *SoCs* se caracterizan por utilizar CPUs de arquitectura ARM, en su mayoría ARMv7 y GPUs de PowerVR de la línea SGX.

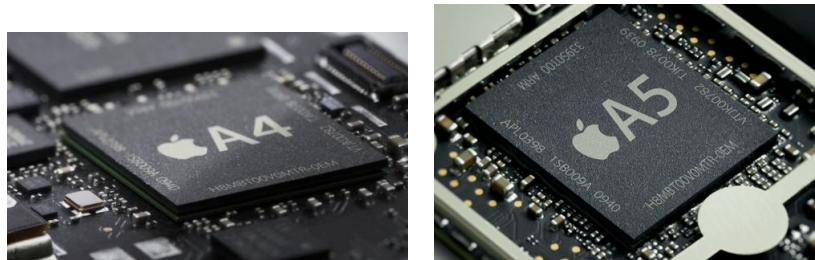


Figura 1.1: *System on Chip*: SoC.

La arquitectura ARM incorpora algunas características de la arquitectura RISC (*Reduced Instruction Set Computer*) como el hecho que las operaciones sean llevadas a cabo sobre un conjunto de registros a tales efectos y no sobre la memoria directamente. Otra característica que tiene de RISC es que tiene un modo simple de direccionamiento donde las direcciones son también guardadas sobre registros destinados a tales efectos. La mayoría de los procesadores están hechos con un ancho de palabra de 32-bit salvo el reciente ARMv8 que incorpora la posibilidad de utilizar ambos anchos de palabra: 32-bit o 64-bit. Otra característica a destacar sobre estos procesadores es que es posible programar sobre ellos utilizando lenguaje C/C++. Por más información sobre esta arquitectura referirse a la web de la empresa <http://www.arm.com>.

Por su parte las GPU utilizadas en los *SoCs* de la serie Apple Ax, son GPUs de PowerVR, una división de la firma Imagination Technologies (<http://www.imgtec.com>) que desarrolla hardware y software para *rendering* 2D y 3D, procesamiento de imágenes y codificación. La función que tiene la GPU es asignar a cada pixel de la pantalla su color para cada cuadro. En particular, estas GPU implementan un concepto innovador de *renderizado* que mejora notoriamente la performance de los gráficos: *Tile-Bassed Deferred Rendering* (TBDR). Este concepto aprovecha la independencia de áreas alejadas de la pantalla y de la correlación de píxeles cercanos y divide la pantalla en *tiles* o baldosas. A cada *tile* se le asocia un procesamiento paralelo y con esto se mejora notablemente la performance respecto al método tradicional: Immediate mode renderers (IMRs), que procesa la pantalla completa. Las imágenes *renderizadas* están hechas por triángulos (polígonos), por lo que uno de los indicadores fundamentales para evaluar la performance de una GPU es la cantidad de triángulos (polígonos) que es capaz de procesar por segundo. En (referencia benchmark: <http://www.anandtech.com/show/4216/apple-ipad-2-gpu-performance-explored-powervr-sgx543mp2-benchmarked>) se puede ver un análisis interesante que compara entre otros, a los dos tipos de GPU que se presentaron en la Tabla 1.1: SGX535 y SGX543MP2. En dicho análisis se muestra por ejemplo que en el mejor de los casos la GPU SGX535 fue capaz de procesar 8.69 millones de triángulos por segundo frente a los 29 millones procesados por la GPU SGX543MP2.

1.4. Software de Apple Inc.

1.4.1. Sistemas Operativos

Para poder desarrollar aplicaciones sobre dispositivos móviles de la firma Apple Inc. es necesario contar con computadoras que corran el sistema operativo **Mac OS X**. Esto puede ser llevado a cabo, ya sea adquiriendo plataformas de desarrollo de la mencionada firma o creando máquinas virtuales que corran dicho sistema operativo. Para la segunda opción (la más económica pero con ciertas dificultades de performance), es necesario que la computadora cuente con virtualización de hardware. Se comenzó trabajando de esta manera hasta el momento de adquirir plataformas de desarrollo que contaran con Mac OS X en forma nativa.

Mac OS X refiere a la versión número 10 (en números romanos) de una serie de sistemas operativos que comenzaron a desarrollarse en la década de los 80 (Mac OS 1 data del año 1984). En los últimos 28 años se han ido sucediendo nuevas versiones que han ido mejorando características en la estructura de datos con la incorporación de la jerarquía de archivos en Mac OS 3 por ejemplo, en la búsqueda de archivos, con la simultaneidad de tareas, multiplicidad de usuarios o incluso con el énfasis en la interfaz de usuario por mencionar algunas características importantes en la evolución de esta familia de sistemas operativos. Dentro de Mac OS X existen distintas versiones, siendo la más actual la Versión 10.8: Mountain Lion lanzada durante 2012.

Por su parte todas las plataformas móviles de Apple Inc corren otro dispositivo de código cerrado: **iOS**. Originalmente llamado así por ser el sistema operativo utilizado por la plataforma iPhone, este sistema operativo está también en las plataformas iPad, iPod Touch y Apple TV en todas sus versiones. La versión más reciente de este SO es el iOS 6.1.

Una de las grandes innovaciones de estas plataformas es el hecho de poder desarrollar aplicaciones y correrlas en el propio dispositivo (por supuesto también sucede lo mismo en el mundo Android). Para poder lograr esto, es necesario como se ha dicho, contar con una máquina que corra Mac OS X y contar con el SDK apropiado llamado **Xcode**. Este entorno de desarrollo y su lenguaje se explican en la sección 1.4.2.

1.4.2. Objective-C

El lenguaje que fue elegido por Apple Inc para desarrollar sobre plataformas móviles es Objective-C. Este lenguaje fue desarrollado en la década de 1980 como un superconjunto de C orientado a objetos. Es decir que es una extensión del standard ANSI C que incorpora un modelo orientado a objetos basado en **Smalltalk**. Una de las diferencias sustanciales del modelo orientado a objetos de Objective-C respecto a otros lenguajes como Java o C++, es el hecho de la invocación de los métodos (procedimientos) de las instancias de clases. En objective-C esta invocación se da enviando mensajes, algo que se hereda de Smalltalk. Así entonces para invocar un método se procede con el siguiente código por ejemplo:

```
[receiver message];
```

Donde *receiver* es un objeto que recibe un mensaje (acción) *message* a realizar. Esta acción puede tener parámetros asociados, como por ejemplo el siguiente código real:

```
[myRectangle setWidth:20.0];
```

Esta diferencia conceptual de utilizar mensajes se representa en el hecho de que en tiempo de compilación estos mensajes no son más que una etiqueta y no están asociados al bloque de código como es el caso de Java o C++. Entonces es factible que suceda el hecho de que ese mensaje o método no

esté implementado por esa clase y recién en tiempo de ejecución es que saltará el error al sustituirse esa etiqueta por un código inexistente, pues un objeto recibe un mensaje para realizar un método que no está dentro de su repertorio. Para esto es que en la documentación de Apple Inc se recomienda utilizar ciertos trucos para garantizar que el objeto que reciba el mensaje sea capaz de responder correctamente, como por ejemplo consultando primero si es capaz de realizar dicha acción y luego en caso de poder realizar dicha acción.

Otro detalle a destacar es que este lenguaje, al igual que Java también soporta la herencia múltiple. Esto es, dado un conjunto de métodos que son comunes a un conjunto de clases pero que no llegan a tener un lazo tan fuerte como para estar jerárquicamente relacionadas con una superclase común, se puede generar una clase abstracta cuyos métodos sean implementados por más de una clase sin necesidad de generar ese vínculo fuerte que es la herencia. Así como en Java existen las interfaces, que hacen esto posible, en Objective-C existen los protocolos. Existen protocolos formales e informales y con métodos obligatorios de implementar y otros opcionales. Una clase que implemente un protocolo dado tiene que tener dentro de su encabezado declarado el nombre del protocolo. Esto es:

```
@interface ClassName : ItsSuperclass < protocol list >
```

Hay otras particularidades del lenguaje pero que no van más allá de la sintaxis como los métodos de clase y los métodos de instancia, como los métodos *get* y *set* para acceder y setear atributos (propiedades) de los objetos, como la notación de *import* en lugar de *include* para quienes están acostumbrados a C y así varias detalles más. Sin embargo más allá de estas y otras diferencias y particularidades resulta un lenguaje relativamente ágil y dentro de todo sencillo de aprender para quien tiene ya un conocimiento de otros lenguajes orientados a objetos.

1.4.3. Xcode: Herramientas y Librerías

Como se dijo anteriormente el entorno de desarrollo de aplicaciones típico es Xcode, el cual es gratuito y permite compilar código C, C++, Objective-C, Objective-C++, Java y AppleScript. Xcode integra en una sola interfaz todo lo que involucra código, diseño de interfaz de usuario (**Interface Builder**) y *debugging*. También viene con un conjunto herramientas útiles para evaluar la performance de la aplicación en distintos aspectos que se llama **Instruments**. Por otra parte viene con un conjunto importante de *Frameworks* entre los cuales se encuentran **Cocoa** y **Cocoa Touch** que proveen de herramientas útiles para desarrollar más fácilmente aplicaciones para Mac OS X e iOS respectivamente.

Las aplicaciones que corren sobre los distintos dispositivos como iPhone, iPod Touch, iPad o AppleTV están desarrolladas en Objective-C pero sobre la base de estas librerías o *Frameworks* de iOS que se pueden separar en cuatro grandes capas según el nivel de abstracción: Cocoa Touch, Media, Core Services y Core OS. Así entonces, dentro de cada capa existen distintos *Frameworks* según la

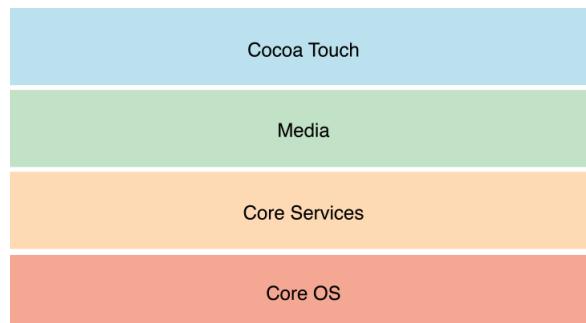


Figura 1.2: Capas de iOS

funcionalidad. A continuación se explica un poco más en detalle el rol de cada capa, los distintos *Frameworks* que tiene cada una y para qué sirven.

1.4.3.1. Cocoa Touch Layer

Cocoa Touch es la capa de más alto nivel de iOS y es la encargada de proveer al desarrollador de ciertos *Frameworks* que permitan lograr distintas tecnologías como la posibilidad de multitarea, el ingreso de órdenes a la aplicación a través de la pantalla táctil, notificaciones y alertas, preservación del estado de la aplicación al salir de la misma, reconocimiento de gestos en la pantalla y otro tipo de funcionalidades de alto nivel. Permiten al desarrollador, sin tener que involucrarse demasiado a bajo nivel, el acceso a determinados servicios que ya están resueltos en forma bastante modular. Cocoa Touch está basado en la arquitectura **Modelo-Vista-Controlador**, en el que se separa en tres áreas distintas el modelo de la información, la interfaz de usuario y el conjunto de reglas que negocian la presentación de la información en base a la interacción con el usuario. Así pues, el usuario y una aplicación se podrían considerar dos sistemas que interactúan. Por su parte el usuario tiene como entrada la vista de la aplicación y como salida tiene su respuesta a esta entrada, generando efectos sobre el control de la aplicación. Por otro lado la aplicación tiene como entrada las órdenes dadas por el usuario que tienen efectos sobre el modelo de la información y este sobre la vista, quien resulta ser la salida de la aplicación. Esta interacción se puede ilustrar con la figura 1.3. Como se

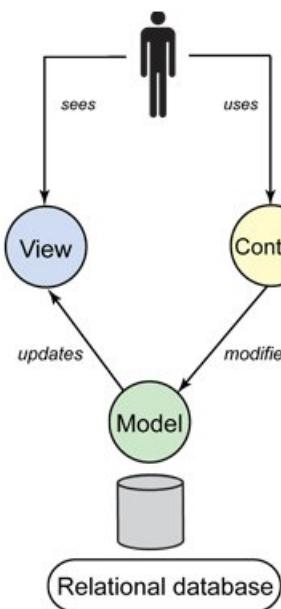


Figura 1.3: Interacción entre las tres partes del MVC

dijo, dentro de Cocoa Touch, existen distintos *Frameworks* enfocados en permitirle al desarrollador resolver en alto nivel distintos aspectos. Los mismos son los siguientes:

- (1) Address Book UI Framework
- (2) Event Kit UI Framework
- (3) Game Kit Framework
- (4) iAd Framework
- (5) Map Kit Framework

(6) Message UI Framework

(7) Twitter Framework

(8) UIKit Framework

Quizá sea bueno mencionar que varias de estas API no fueron utilizadas en el presente proyecto dada su función específica y que no fueron necesarias. Sin embargo hay una en particular que tiene bastante importancia y que permite la mayoría de las funcionalidades básicas que toda aplicación tiene. Se trata del **UIKit**, encargado de gestionar la aplicación, su interfaz de usuario y gráficos, encargado soportar eventos frente al toque de la pantalla, de manejar sensores como el acelerómetro y giroscopio, y de tener acceso a la cámara y galería de fotos entre lo más importante a destacar. El soporte de la multitarea y de **Storyboards** también está a cargo de este *Framework*.

Hay funcionalidades que han ido cambiando con las distintas versiones de iOS. Una de ellas y quizás una que ha generado bastantes diferencias respecto a versiones anteriores a iOS 5, es esta última, el Storyboard, una herramienta muy útil de programación gráfica, que permite generar instancias de clases y vínculos entre las mismas en forma visual a la vez de ser una contraparte de interfaz de usuario. Con una biblioteca de objetos disponibles, listos para ser usados, mediante el uso de Storyboard se hace accesible con algunas horas de dedicación implementar aplicaciones sencillas. Esta herramienta vino para sustituir los archivos *.nib* que permitían diseñar la interfaz pero no tantas funcionalidades programáticas como el Storyboard. En particular éste último permite agregar la funcionalidad de *segues*, encargados de vincular un *ViewController* con otro. Este tipo de diferencias vinieron con la idea de evitar la necesidad de implementar ciertos bloques de código en forma repetitiva. Un Storyboard luce como en la figura 11.2.

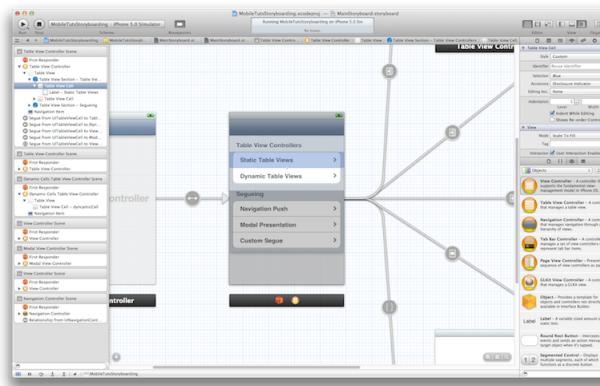


Figura 1.4: Ejemplo de Storyboard.

Si bien se podría extender bastante más la explicación sobre los detalles de Cocoa Touch, a los efectos del presente proyecto, no es de tanta relevancia excederse en este punto.

1.4.3.2. Media Layer

Media Layer es la capa encargada de gestionar correctamente elementos multimedia y es posible distinguir tres grandes grupos que engloban distintos *Frameworks*: Gráficos, Audio y Video.

Dentro de las tecnologías más destacadas está todo lo vinculado a **gráficos** 2D y 3D, dentro de lo que se puede incluir algunos *Frameworks* bastante utilizados en el presente proyecto, tales como: **Core Graphics**, muy utilizado para dibujos 2D, **Quartz Core**, quien contiene las herramientas necesarias para interactuar con otro *Framework* para animación de vistas, de una capa de más bajo nivel como *Core Animation*, que es comentado más adelante en la sección 1.4.3.3. También es parte

de lo vinculado a gráficos, el **Framework Core Image**, contenido lo vinculado a procesamiento de imágenes a través filtros que utilizan directamente la unidad de procesamiento de gráficos (GPU) y otros dos *Frameworks* bastante importantes en lo que respecta a *rendering* como **OpenGL ES** y **GLKit** (utilizado por el motor de juegos *Isgl3d* entre otros).

Por otra parte hay otra gran familia de *Frameworks* dentro de Media Layer que apunta a resolver todo lo vinculado al manejo de audio, ya sea de grabación como procesamiento y reproducción de alta calidad. Existen algunos SDK como **iSpeech** o **Dragon Mobile** que resuelven de manera similar al proyecto Siri, el procesamiento de la voz humana reconociendo palabras e interpretando, que utilizan algunos de los *Frameworks* de procesamiento de audio de esta familia.

En cuanto a lo vinculado al manejo de video, como parte de esta capa, se tienen dos *Framework* importantes con distintos niveles de abstracción: **MediaPlayer** y **AVFoundation**. También existen otros *Frameworks* fuera de esta capa que son capaces de manejar video como la clase `UIImagePickerController` (muy utilizada en el proyecto) del mencionado `UIKit`. En la figura 1.5 se esquematiza el nivel de abstracción de los *Frameworks* de las distintas capas que son capaces de manejar multimedia. Con `MediaPlayer` es posible reproducir audio y video muy fácilmente en determinada área

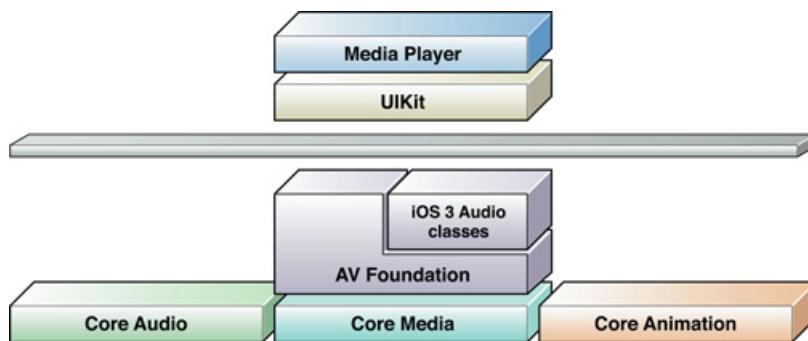


Figura 1.5: Frameworks de las distintas capas para manejo de video

de la pantalla ya sea desde un URL o de un archivo, es posible mostrar o no los elementos de control del video así como también controlar, volumen y tamaño de la pantalla. Por su parte, con AVFoundation es posible capturar con la cámara, reproducir, editar y procesar audio y video. Es posible implementar ciertos protocolos que hace de esto algo relativamente sencillo.

1.4.3.3. Core Services

Core Services es la capa de más bajo nivel de iOS y contiene los elementos fundamentales sobre los que se construyen las capas superiores. Es posible que al comenzar a programar para iOS no se tenga mucha interacción con esta capa pero sin embargo existen algunos conceptos importantes de esta capa que sí vale la pena mencionar dado que en el presente proyecto se tuvieron que entender y discutir. Una de ellas es la *Automatic Reference Counting* o **ARC**. Esta funcionalidad compete a la reserva y liberación de memoria por parte de los objetos. La idea básica es lograr que el uso de memoria sea el mínimo posible, logrando que los objetos existan en la medida que son necesarios y que su memoria sea liberada ni bien sea posible. Típicamente, al crear una instancia de un objeto se incrementa un contador y al liberar se decrementa y entonces se tiene cierto control sobre la reserva y liberación de memoria en base al contador. Sin embargo, la liberación de memoria reservada por objetos queda bajo la responsabilidad del desarrollador y en casos de un código complejo puede llegar a ser habitual olvidarse de la liberación de memoria. Lo anterior refiere a una gestión manual de la reserva y liberación conocida como *manual retain-release*. Para no tener que enfrentar este tema y poder instanciar clases sin tener presente la posterior liberación de memoria (pues quizás se sepa cuándo no será más necesario un objeto o no), se utiliza ARC. Esta funcionalidad evalúa el ciclo

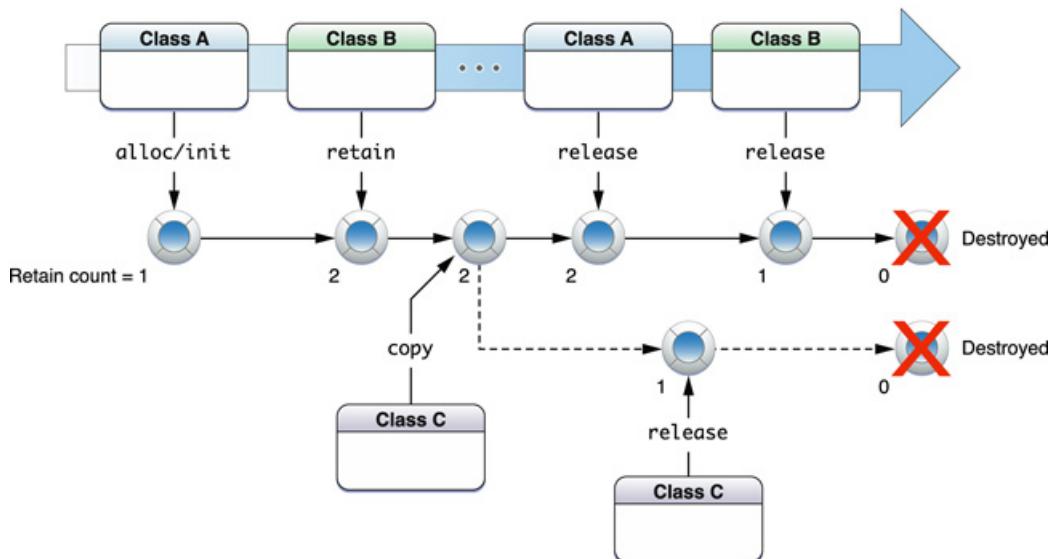


Figura 1.6: Ciclo de vida de objetos, Manual-Retain-Release.

de vida de los objetos y agrega código en tiempo de compilación en caso de considerarlo necesario. Es bueno aclarar que esto refiere a memoria reservada pura y exclusivamente por objetos, es decir mediante *alloc*. En caso de tratarse de memoria reservada para variables de lenguaje C (*malloc*), es necesario proceder de igual manera que en dicho lenguaje, liberando la memoria mediante un *free*. Además del ARC, Core Services permite el manejo de archivos XML y manejo de base de datos SQL así como también la protección de datos cuando el dispositivo está bloqueado entre otros servicios importantes. Tiene varios *Frameworks* como **Core Media** que logran un nivel aún más bajo que AVFoundation para el manejo de multimedia, **Quick Look** para las vistas previas de archivos, **Social** que viene a suplantar el *Framework* para la utilización de Twitter que existe en iOS 5 y extiende la gestión para otras redes sociales, **Core Motion** para el manejo de sensores como el acelerómetro y el giroscopio, **Core Telephony** para el manejo de la información de red del dispositivo como elemento de la red de telefonía, **CFNetwork** para el manejo de protocolos de red como http, https, ftp y resolución de servidores DNS, entre otros *Frameworks* importantes dentro de la capa.

1.4.3.4. Core OS

Con esta capa de iOS en general es difícil que el desarrollador tenga que involucrarse directamente dado que es la de más bajo nivel. Salvo que se esté frente a una aplicación que requiera aspectos de seguridad o comunicación con HW externo, esta capa solamente existe para ser la base sobre la cual se desarrollan los *Frameworks* de las capas de más alto nivel. Los distintos *Frameworks* que tiene están enfocados en resolver temas de procesamiento basados en el hardware de iOS, en comunicarse con dispositivos externos basados en iOS y de garantizar la seguridad de los datos de una aplicación.

1.4.3.5. Simulador

Uno de los detalles más importantes del entorno de desarrollo es la capacidad de simular lo que se programa antes de probarlo en un dispositivo. Esto es útil por cuestiones de seguridad e incluso permite programar sin la necesidad de contar con una plataforma. Esto existe para *Xcode* y es necesario decir que funciona muy bien, generando una representación bastante fiel de lo que sucede en el dispositivo real. La única crítica que se le podría hacer es el hecho de no contar con cámara y

para el caso de aplicaciones de realidad aumentada esto es algo bastante importante. Sin embargo, sin ser eso, el simulador cuenta con conexión a internet, pantalla multitáctil, con información de GPS ingresada por el programador, acceso a la galería de fotos, capacidad de procesamiento y todas las funcionalidades que un dispositivo real tiene.

1.4.3.6. Instruments

Dentro de las herramientas que vienen con el entorno de desarrollo viene *Instruments*, un conjunto de herramientas que permiten analizar la performance de una aplicación para iOS o para Mac OS X desde distintos puntos de vista. Resulta muy útil pues muchas veces sucede que una aplicación compila y se ejecuta correctamente y sin embargo puede que el desarrollador no esté conforme en cuanto a los tiempos de procesamiento o el uso de memoria consumido.

Para poder hacer uso del *Instruments*, es necesario correr la aplicación en modo *Profile*. Eso despliega una ventana como la de la Figura 1.7. Allí es posible elegir dentro de cada una de las posibilidades que ofrece *Instruments*, si se quiere analizar tiempos, memoria, recursos de CPU, *multithreading* entre otros tipos de datos de interés que es posible recoger de la aplicación. También es posible elegir la plataforma, ya sea iOS, simulador iOS o Mac OS X.



Figura 1.7: Distintas opciones de *Instruments*.

Luego de elegir el tipo de datos a ser recolectados según lo que se busque analizar, se despliega una ventana como se ve en la Figura 1.8. Allí es necesario registrar durante varios segundos los datos mientras se corre la aplicación y luego de terminado el registro, *Instruments* dedica un tiempo a analizar los datos recolectados. En el detalle inferior, se desglozan los procesos que corre la aplicación en forma de árbol. Cuando se desea medir tiempos por ejemplo, esto resulta muy útil porque entre otras cosas se puede analizar qué porcentaje del tiempo de la aplicación es consumido por un método en particular. Esto es posible simplemente buscando dentro del árbol mencionado, al método y leyendo el valor asignado de tiempo. Para el caso de análisis de memoria también es posible identificar fácilmente en qué parte del código se está dando algún problema de reserva de memoria no liberada. En el presente proyecto se hizo uso principalmente del **Time Profiler** que permite analizar tiempos y del **Memory Leak** que permite hacer un análisis de la reserva de memoria.

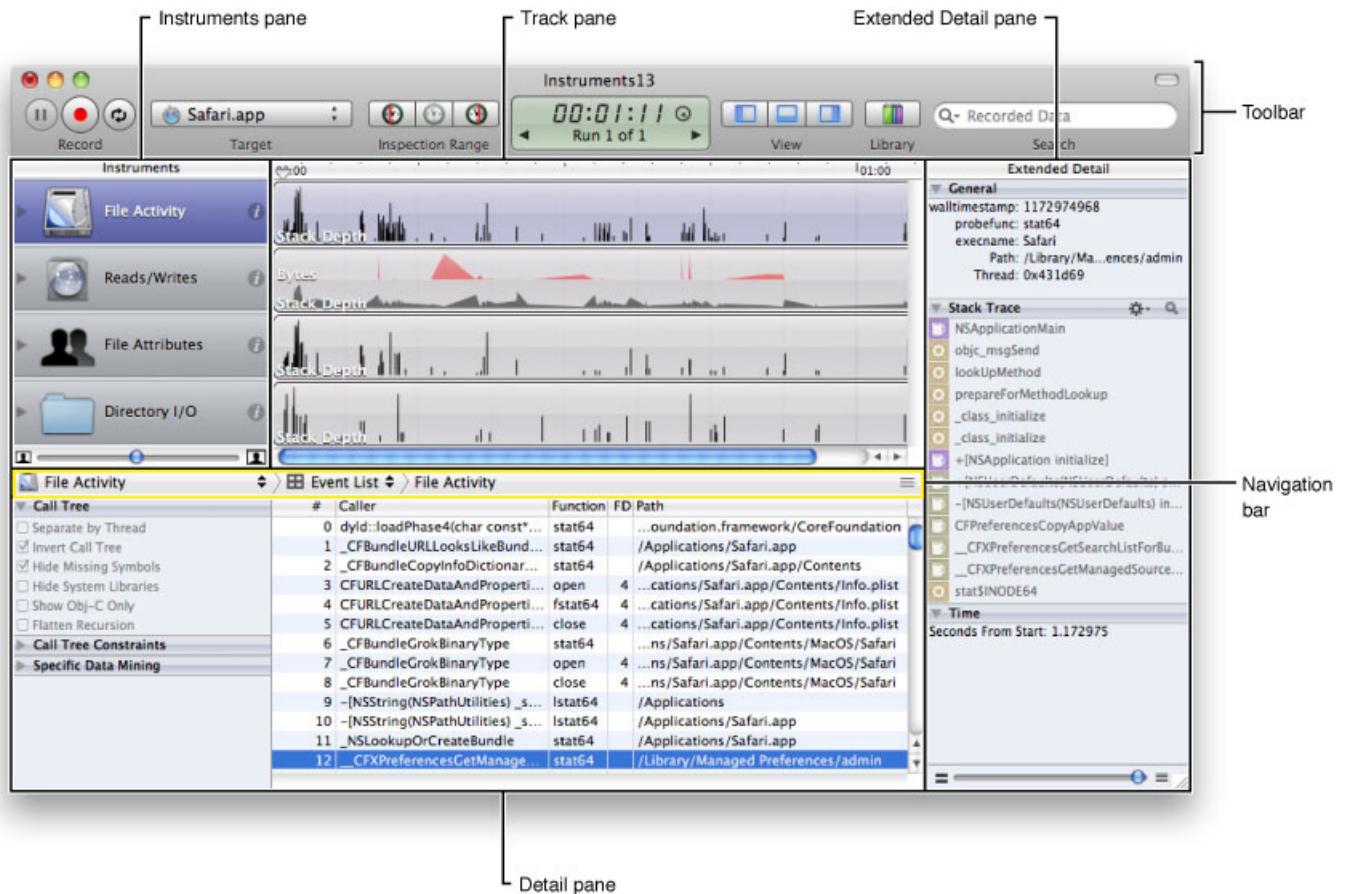


Figura 1.8: Trazado y análisis de datos recogidos.

no liberada. Con los mismos fue posible optimizar tiempos en determinados métodos del procesamiento, así como también eliminar problemas de memoria no liberada que desencadenaban en la interrupción abrupta de la aplicación luego de llegar a un cierto nivel de reserva. Esta interrupción abrupta es una forma de proteger la memoria del dispositivo y evitar que se vea afectada cierta memoria útil a otros efectos. El resto de las herramientas de *Instruments* fueron probadas pero no utilizadas en detalle para resolver problemas particulares.

1.5. Herramientas

Herramientas

1.5.1. GIT

1.5.2. GoogleCode

1.5.3. Github

CAPÍTULO 2

Identificación

2.1. Introducción

Introducción

2.2. Técnicas de identificación

Técnicas de identificación

2.3. QR

Ventajas del Lenguaje C para procesamiento de imágenes

2.3.1. Identificadores QR: una realidad cotidiana.

El uso de los identificadores QR (Quick Response), es cada vez más generalizado. Últimamente, debido al incremento significativo del uso de *smart devices*, el hecho de poder contar con una cámara, cierto poder de procesamiento y por lo general hasta una conexión móvil a internet, hace que sea cada vez más frecuente encontrar aplicaciones con el poder de reconocer QRs. Comenzaron utilizándose en la industria automovilística japonesa como una solución para el trazado en la línea de producción, pero su campo de aplicación se ha diversificado y hoy en día se pueden encontrar también como identificadores de entradas deportivas, tickets de avión, localización geográfica, vínculos a páginas web y en algunos casos también como tarjetas personales.

2.3.2. ¿Qué son los QR?

Los QRs son una extensión de los códigos de barras. Incorporan una segunda dimensión lo cual es una gran ventaja ya que pueden almacenar mucho más información. Existen distintos tipos de QR, con distintas capacidades de almacenamiento que dependen de la versión, el tipo de datos almacenados y del tipo de corrección de errores. En su versión 40 con detección de errores de nivel L, se pueden almacenar alrededor de 4300 caracteres alfanuméricos o 7000 dígitos (frente a los 20-30 dígitos del código de barras) lo cual lo hace muy flexible para cualquier tipo de aplicación de identificación.

En la Figura 11.9 se pueden ver las distintas partes que componen un QR, como por ejemplo el bloque de control, compuesto por las tres esquinas idénticas que dan información de la posición, la información de alineamiento y el patrón de sincronismo; así como también la indicación de versión, formato y la corrección de errores. Fuera de toda esa información, que podría verse como el encabezado, haciendo analogía con los paquetes de las redes de datos, se encuentran los datos propiamente dicho, que podrían verse como el cuerpo del paquete.

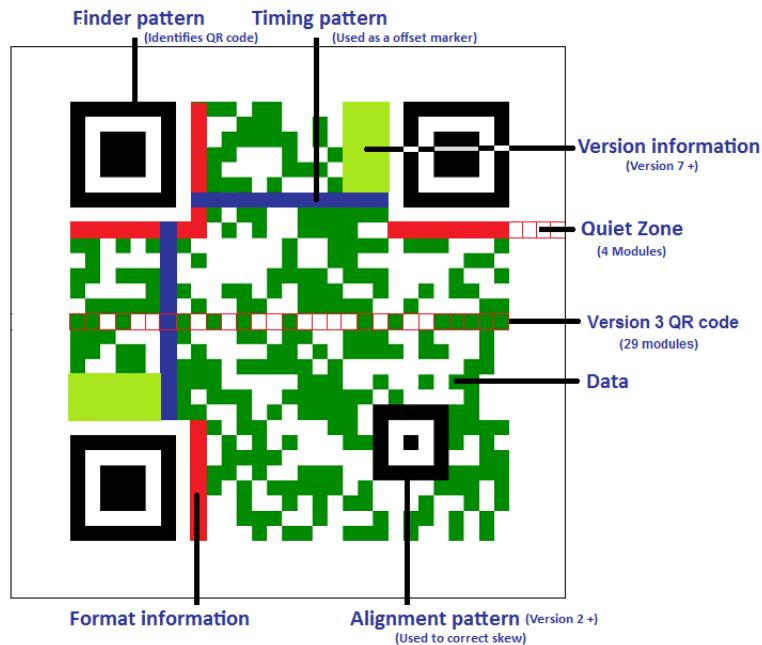


Figura 2.1: Las distintas componentes de un QR. Fuente (poner fuente).

2.3.3. Codificación y decodificación de códigos QR.

Es fácil darse cuenta que la codificación resulta mucho más sencilla que la decodificación. Para la codificación es necesario comprender el protocolo, las distintas variantes y el tipo de información que se pretende almacenar. Sin embargo, para la decodificación, además de tener que cumplir con lo anterior, es necesario contar con buenos sensores y ciertas condiciones de luminosidad y distancia que favorezcan a la cámara y se traduzcan en buenos resultados luego de la detección de errores. Si bien la plataforma es importante para lograr buenos resultados, dada una plataforma, existen variadas aplicaciones tanto para iOS como para Android que cuentan con performances bastante diferentes en función del algoritmo de procesamiento utilizado.

Debido a que el centro del presente proyecto no fue la codificación y decodificación de QRs, y que además ya existen distintas librerías que resuelven muy bien este problema, se optó por investigar varias de ellas e incorporar la más adecuada a la aplicación.

Entre todas las librerías que resuelven la decodificación, las llamadas ZXing y ZBar son quizás las más destacadas, por su popularidad, simplicidad y buena documentación para la fácil implementación. ZXing, denominada así por “Zebra Crossin”, es una librería gratis y en código abierto desarrollada en java y que tiene implementaciones que están adaptadas para otros lenguajes como

C++, Objetive-C y JRuby, entre otros.

Por su parte ZBar también tiene soporte sobre varios lenguajes y cuenta con un kit de desarrollo interesante para lograr fácilmente aplicaciones que integren el lector de QR. Se trabajó sobre el código de ejemplo que contiene la implementación de las clases principales para obtener un lector y finalmente se optó por utilizar esta librería para los fines de la aplicación. El lector del código de ejemplo consta de una clase *ReaderSampleViewController* que hereda de *UIViewController* y que implementa un protocolo llamado *ZBarReaderDelegate*. Al presionarse el botón de detección se crea una instancia de la clase *ReaderSampleViewController* y se presenta la vista previa de la cámara. Luego el protocolo se encarga de la captura y procesamiento del QR almacenando como resultado la información embebida en este en la variable denominada *ZBarReaderControllerResults*. Esta variable luego se mapea en una *hash table* con el contenido en formato *NSDictionary*. De esta manera se accede fácilmente al contenido en formato legible y es fácil de hacer una lógica de comparación y búsqueda en una base de datos.

2.3.4. Expresiones artísticas con QRs.

La opción de usar los QR de una manera distinta ha comenzado a ser notoria en los últimos tiempos. Hay quienes desafían a la información *cruda de 1s y 0s* incorporando imágenes y modificando colores y contornos en los QR tradicionales para lograr un valor estético además del funcional. Véase en la figura 11.10 un ejemplo de cómo puede lograrse el mismo resultado pero con el valor agregado de originalidad.



Figura 2.2: Ejemplo de un QR creativo. Fuente (poner fuente).

2.4. SIFT

El algoritmo SIFT, acrónimo de “Scale Invariant Feature Transform”, es un algoritmo de visión artificial [?], [?], que se encarga de extraer características distintivas de las imágenes en escala de grises. Mediante estas, es posible luego reconocer dicha imagen dentro de una base de datos o incluso dentro de otra imagen mayor con otra cantidad de elementos en desorden. Estas características

son invariantes a factores de escala, traslación, rotación y parcialmente invariantes a cambios de iluminación y afinidades. El algoritmo consta básicamente de cuatro pasos que se explicarán brevemente en secciones subsiguientes.

2.4.1. Detección de extremos en el espacio-escala.

Se busca encontrar dentro del *scale-space* (espacio-escala) de la imagen puntos característicos; invariantes a la traslación, el escalado y la rotación de la misma. Además esos puntos deben ser mínimamente afectados por el ruido y pequeñas distorsiones. Serán los puntos extremos (máximos o mínimos) obtenidos de las diferencias Gaussianas aplicadas al *scale-space* de la imagen. El *scale-space* de una imagen se define como una familia de imágenes $L(x, y, \sigma)$ que se obtienen de convolucionar un núcleo Gaussiano variable en su desviación estándar $G(x, y, \sigma)$ con una imagen de entrada $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

donde $*$ denota la convolución en x e y , y además:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Una imagen diferencia de Gaussianas, $D(x, y, \sigma)$, se define entonces de la siguiente manera:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma)$$

con k un factor multiplicativo constante.

Dado un valor inicial para σ , se le realiza a la imagen un número s de diferencias Gaussianas con la desviación estándar variando de manera creciente a lo largo de una octava (hasta obtener un $\sigma' = 2\sigma$). Para obtener s intervalos enteros dentro de dicha octava el valor de k en cada diferencia Gaussiana debe ser de $2^{\frac{1}{s}}$. Una vez calculadas las diferencias Gaussianas a lo largo de la octava, la imagen se submuestra tomando 1 de cada 2 píxeles en filas y columnas y se procede de la misma manera. La cantidad de octavas involucradas en el cálculo así como la cantidad de diferencias Gaussianas calculadas por octava son un parámetro a determinar. En la figura 2.3 se puede ver lo anterior explicado gráficamente.

Una vez que se obtiene la “pirámide” de diferencias Gaussianas anterior, se buscan para cada “piso” de la misma extremos locales quienes se transformarán en candidatos a puntos clave. Para una $D(x, y, \sigma)$ determinada y en una octava determinada, un punto (x_0, y_0) será un máximo (mínimo) relativo si es mayor (menor) a sus 8 puntos vecinos dentro de su nivel y a sus 9 puntos vecinos de cada uno de los niveles inferior y superior. Si el punto se encuentra en una $D(x, y, \sigma)$ de transición entre 2 octavas, se buscan los puntos vecinos correspondientes del nivel superior (inferior). Ver figura 2.4.

2.4.2. Localización exacta de puntos clave.

La búsqueda de extremos en las diferencias de Gaussianas produce múltiples candidatos entre los que se encuentran puntos con poco contraste; los cuales no son estables a cambios de iluminación y al ruido. Para quitarlos se procede de la siguiente manera.

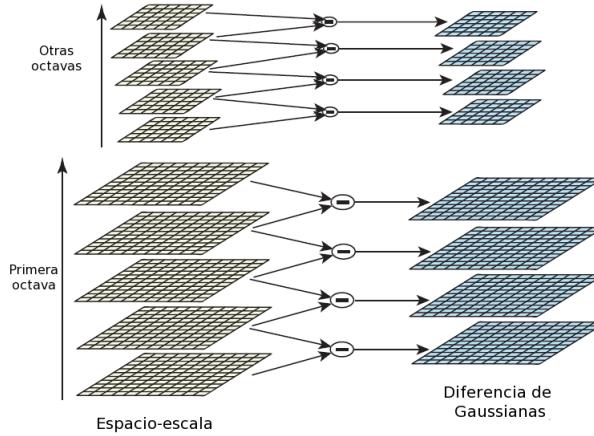


Figura 2.3: Para cada octava, la imagen original es convolucionada repetidamente con Gaussianas de desviación estándar variable para producir el *scale-space* de la izquierda. Imágenes adyacentes del *space-scale* son restadas para lograr la diferencia de Gaussianas de la derecha. Después de cada octava, la imagen borrosa es submuestreada por un factor de dos y el proceso se repite. Figura tomada de [?].

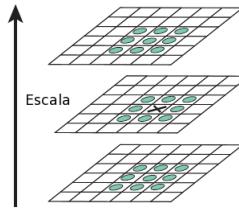


Figura 2.4: Máximos y mínimos de las imágenes diferencia de Gaussianas son obtenidos comparando cada píxeles con sus vecinos en la misma escala, y en las escalas adyacentes. Figura tomada de [?].

Primero se realiza una expansión de Taylor de grado 2 entorno a cada extremo detectado (x_0, y_0, σ_0) :

$$D(\chi) = D + \frac{\partial D^T}{\partial \chi} \chi + \frac{1}{2} \chi^T \frac{\partial^2 D}{\partial \chi^2} \chi \quad (2.1)$$

donde D y sus derivadas son evaluadas siempre en el punto en cuestión y $\chi = (x, y, \sigma)^T$ es la posición relativa al mismo. Derivando la aproximación anterior e igualándola a cero se obtiene:

$$\bar{\chi} = -\frac{\partial^2 D^{-1}}{\partial \chi^2} \frac{\partial D}{\partial \chi} \quad (2.2)$$

Reemplazando (2.2) en (2.1) se obtiene el valor del máximo local:

$$D(\bar{\chi}) = D + \frac{1}{2} \frac{\partial D^T}{\partial \chi} \bar{\chi}$$

Finalmente, si $|D(\bar{\chi})| < 0,03$ el punto es eliminado de la lista de puntos clave; suponiendo que D toma valores entre 0 y 1.

Además de quitar aquellos puntos con poco contraste, hay que quitar a los puntos candidatos que pertenecen a una línea y no a una esquina. Para ello, sea H la matriz Hessiana de $D(x, y, \sigma)$

evaluada en un punto extremo de las diferencias de Gaussianas determinado (x_0, y_0, σ_0) , se estará en presencia de un borde (línea) si sus valores propios α y β son uno grande y el otro pequeño. Lo anterior es equivalente a trabajar con los siguientes resultados:

$$\begin{aligned} \text{Trazza}(H) &= \frac{\partial^2 D}{\partial x^2} + \frac{\partial^2 D}{\partial y^2} = \alpha + \beta \\ \text{Det}(H) &= \frac{\partial^2 D}{\partial x^2} \times \frac{\partial^2 D}{\partial y^2} - \left(\frac{\partial^2 D}{\partial x \cdot \partial y} \right)^2 = \alpha \cdot \beta \end{aligned}$$

Sea la $\alpha = r \cdot \beta$, la condición se reduce a:

$$\frac{\text{Trazza}(H)^2}{\text{Det}(H)} < \frac{(r+1)^2}{r}$$

Luego de varios experimentos, se propone un umbral de $r = 10$. Véase que conforme aumenta la relación r entre ambos valores propios también lo hace la relación entre el cuadrado de la traza de la matriz Hessiana y su determinante.

2.4.3. Asignación de orientación.

Mediante la asignación de una orientación a cada punto de la imagen basada en características locales de la misma, los puntos clave pueden ser descriptos relativos a estas orientaciones y de esta manera lograr características invariantes a las rotaciones. Para cada punto clave obtenido $D(x_0, y_0, \sigma_0)$, se busca su imagen borrosa correspondiente en el espacio escala $L(x, y, \sigma_0)$ y se determina el módulo de su gradiente $m(x, y)$ y la fase del mismo $\theta(x, y)$ utilizando diferencias entre píxeles:

$$\begin{aligned} m(x, y) &= \sqrt{(\Delta L_x)^2 + (\Delta L_y)^2} \\ m(x, y) &= \sqrt{[L(x+1, y) - L(x-1, y)]^2 + [L(x, y+1) - L(x, y-1)]^2} \\ \theta(x, y) &= \tan^{-1} \left(\frac{\Delta L_y}{\Delta L_x} \right) \\ \theta(x, y) &= \tan^{-1} \left(\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)} \right) \end{aligned}$$

Para determinar de una forma fiel la orientación de cada punto clave, ésta es determinada tomando en cuenta las direcciones de todos los puntos de la imagen dentro de cierto entorno al mismo. Se genera entonces un histograma de direcciones con valores que varían de a 10 grados, ponderado por el módulo del gradiente y una ventana Gaussiana circular centrada en el punto clave, de desviación estándar igual a 1.5 veces el valor del nivel del en cuestión. Cada máximo en el histograma corresponde a la dirección dominante en el gradiente local y será la asignada al punto clave. Si existen en el histograma otros máximos secundarios de valor mayor o igual al 80 % del máximo principal, estos serán utilizados para generar nuevos puntos clave con esa dirección. Sólo al 15 % de los puntos clave se les asigna más de una dirección.

2.4.4. Descriptor de puntos clave.

Hasta el momento, se le ha asignado a cada punto clave una escala, una locación y una orientación. El siguiente paso es determinar para cada punto clave un descriptor relativamente invariante a cambios de iluminación y afinidades, basado en el entorno del mismo.

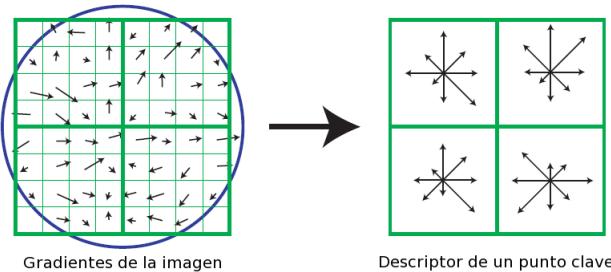


Figura 2.5: Izq.: La ventana Gaussiana pondera los valores de módulo y fase en la vecindad de los puntos de interés. Der.: los histogramas con 8 direcciones posibles realizados para cada subregión. Figura tomada de [?].

Una vez determinadas la magnitud y fase del gradiente entorno a un punto clave, una ventana Gaussiana centrada en este pondera los valores de módulo y fase de 4×4 subregiones cuadradas en la vecindad del mismo, cada una formada por 16 píxeles. Nuevamente se genera para cada subregión un histograma de 8 direcciones distintas. Se obtiene finalmente para cada punto clave un descriptor de $4 \times 4 \times 8 = 128$ valores.

En la figura 2.5 se ve cómo se computan los descriptores para cada punto clave. En el ejemplo se utilizan únicamente $2 \times 2 = 4$ subregiones en vez de $4 \times 4 = 16$.

CAPÍTULO 3

Detección

3.1. Típos de características

3.2. Bordes y esquinas

3.2.1. Detector de bordes de Canny

3.2.2. Detector de bordes y esquinas de Harris

3.2.3. SUSAN Y FAST

3.3. Líneas y segmentos de línea

3.3.1. Detector de líneas de Hough

3.3.2. Detector de segmentos de línea: LSD

3.3.3. Detector de segmentos de línea: EDLines

3.4. Regiones y puntos de interés

3.4.1. FAST

3.4.2. Blobs

3.5. Descriptores

SIFT (puntero a capítulo que tiene SIFT para reconocimiento o mismo acá)
SURF, ETC ETC.

CAPÍTULO 4

Marcadores

La inclusión de *marcadores* en la escena ayuda al problema de extracción de características y por lo tanto al problema de estimación de pose [9]. Estos por construcción son elementos que presentan una detección estable en la imagen para el tipo de característica que se desea extraer así como medidas fácilmente utilizables para la estimación de la pose.

Los marcadores planos se pueden obtener mediante la construcción en una geometría coplanar de una serie de primitivas identificables como esquinas, segmentos o líneas. Un único marcador plano puede contener por si solo todas las seis restricciones espaciales necesarias para definir un marco de coordenadas asociado a su pose.

Como se explica en la sección ?? el problema de estimación de pose requiere de una serie de correspondencias $\mathbf{M}_i \leftrightarrow \mathbf{m}_i$ entre puntos 3D en la escena en coordenadas del mundo y puntos en la imagen.

En el primer lugar se explican brevemente algunos de los sistemas de Realidad Aumentada más populares basados en marcadores planos. En segundo lugar se propone el diseño de un marcador específico para la aplicación a este proyecto y se desarrollan las soluciones a los algoritmos de detección de dicho marcador mostrando algunos resultados parciales en el proceso. Por último se muestran algunos resultados de la detección.

4.1. Sistemas basados en marcadores planos

Existen muchos sistemas de visión basados en *marcadores planos* con aplicación en Realidad Aumentada y Navegación. Algunos de ellos son *ARToolKit* [?], *ARTag* [?] y *ARToolKitPlus* [?] utilizados para Realidad Aumentada. A continuación se realiza una breve descripción del funcionamiento general de los mismos.

Los sistemas basados en marcadores planos utilizan típicamente marcadores bitonales. Esto permite reducir la sensibilidad a las condiciones de luz de la escena y a las configuraciones de la cámara por lo que no hay necesidad de identificar tonos de grises y la regla de decisión para cada píxel puede ser reducida, en la versión más simple, a un umbral o *threshold* [?]. El diseño de los marcadores depende en gran medida de la aplicación. En la figura 4.1 se muestran algunos marcadores planos para aplicaciones de Realidad Aumentada en donde cada uno de ellos provee suficientes puntos para permitir el cálculo de pose tridimensional y adicionalmente contienen cierta información en su interior para permitir su identificación.

Es importante que los marcadores puedan ser localizados en un campo de visión amplio para permitir la correcta detección bajo la distorsión asociada a la transformación proyectiva que lleva el marcador en el mundo real al plano de imagen. Por otro lado, si los marcadores contienen informa-

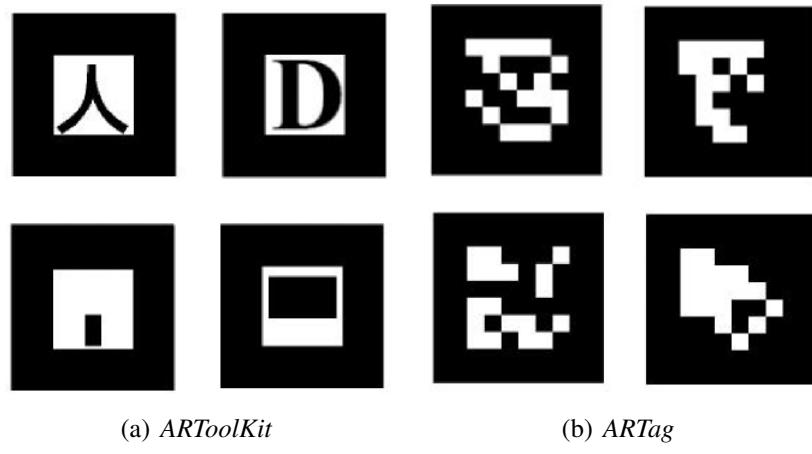


Figura 4.1: Cuatro ejemplos de marcadores para los sistemas de Realidad Aumentada indicados. Figura tomada de [?].

ción en su interior, esta no debe ser muy densa para permitir la recuperación de la misma a mayor distancia. Típicamente, esta información es solo una identificación entre marcadores de un mismo sistema por lo que la información es poca y esto no es un problema.

Estos sistemas contienen ciertos puntos característicos con los que se realiza el cálculo de pose. En general su contorno es basado en un cuadrilátero y se utilizan las cuatro esquinas del contorno del marcador para realizar el cálculo.

Si se utiliza un único marcador la cantidad de puntos necesarios para la estimación de pose resultan ser pocos lo que puede ser una desventaja en cuanto a la precisión del algoritmo de estimación de pose. Esto se puede mejorar construyendo un marcador más complejo compuesto de una serie de marcadores y mediante la identificación de los mismos asignar los puntos correspondientes para la estimación de pose.

4.1.1. ARToolKit

ARToolKit es un muy popular sistema de marcadores planos para Realidad Aumentada e Interacción Hombre-Computador. Su popularidad reside ser de los primeros proyectos en utilizar la Realidad Aumentada en dispositivos móviles y también debido a que es un proyecto de código abierto.

Los marcadores bitonales consisten en un cuadrado con borde negro y un patrón en el interior. La primer etapa del proceso de reconocimiento consisten en detectar los bordes negros. Esto se realiza buscando grupos conexos de píxeles (*blobs*) que están por debajo de un determinado umbral. Posteriormente se extrae el contorno de cada grupo esos grupos que están rodeados por cuatro líneas rectas son marcados como marcadores potenciales. Las cuatro esquinas de cada marcador potencial son utilizados para calcular la homografía y así remover la distorsión perspectiva. Con el marcador en una vista canónica, se procede a identificar el patrón interno muestreando en una grilla, de por lo general 16×16 o 32×32 , los valores de gris. Con esto se construye un vector característico y se compara por correlación con una librería de vectores de característicos logrando la identificación del mismo.

Este sistema tiene algunas desventajas o “puntos débiles”. En primer lugar la detección es basada en un umbral por lo que las condiciones de iluminación pueden afectar fuertemente la efectividad de la misma. Dado que el código esta disponible este se puede modificar para realizar *threshold* local o adaptivo por ejemplo. Otras desventajas están relacionadas con el proceso de identificación

del marcador frente a la librería.

4.1.2. ARTag

ARTag es otro sistema de marcadores planos para Realidad Aumentada e Interacción Hombre-Computador que surge como una evolución de ciertos aspectos de *ARToolKit*. Los marcadores son también bitonales y basados en un borde negro. En contraste con el *ARToolKit* este sistema utiliza un enfoque basado en bordes por lo que es más robusto a condiciones de iluminación. Los bordes son unidos en segmentos que a su vez se unen en cuadriláteros. Al igual que con *ARToolKit* con las esquinas se calcula la homografía y se muestrea en el interior del marcador pero con una grilla de 6×6 .

El sistema puede lidiar con condiciones de iluminación cambiantes, occlusiones y segmentos partidos hasta cierto punto. Otra mejora notable con respecto a *ARToolKit* reside en el sistema de identificación de marcadores. El proceso de identificación de los marcadores entre sí es a su vez más veloz y robusto que el de *ARToolKit*.

4.2. Marcador QR

El enfoque elegido para la detección de características utilizando marcadores parte del trabajo de fin de curso denominado Autoposicionamiento 3D de *Matías Tailanián y Santiago Paternain* para el curso *Tratamiento de imágenes por computadora* de Facultad de Ingeniería, Universidad de la República[?]. La elección se basa principalmente en los buenos resultados obtenidos para dicho trabajo con un enfoque relativamente simple. El trabajo desarrolla, entre otras cosas, un diseño de marcador y un sistema de detección de marcadores basado en el detector de segmentos LSD[7] por su buena *performance*.

El marcador utilizado está basado en la estructura de detección incluida en los códigos *QR* y se muestra en la figura 4.2. Éste consiste en tres grupos idénticos de tres cuadrados concéntricos superpuestos en “capas”. La primer capa contiene el cuadrado negro de mayor tamaño, en la segunda capa se ubica el cuadrado mediano en color blanco y en la última capa un cuadrado negro pequeño. De esta forma se logra un fuerte contraste en los lados de cada uno de los cuadrados facilitando la detección de bordes o líneas. El resultado de una detección de líneas para esta configuración produce para cada cuadrado la detección de sus lados. A diferencia de los códigos *QR* la disposición espacial de los grupos de cuadrados es distinta para evitar ambigüedades en la identificación de los mismos entre sí.

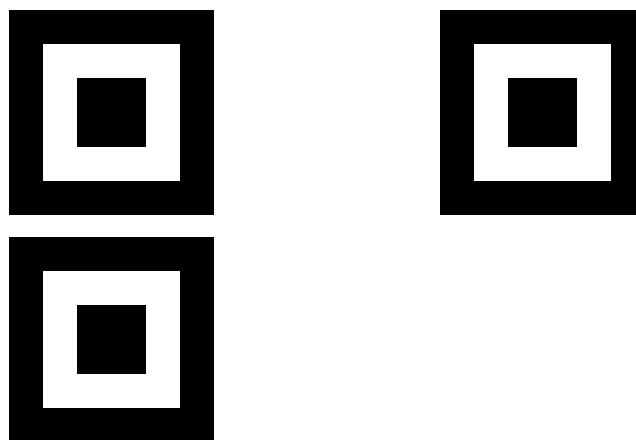


Figura 4.2: Marcador propuesto basado en la estructura de detección de códigos QR.

4.2.1. Estructura del marcador

A continuación se presentan algunas definiciones de las estructuras básicas que componen el marcador propuesto. Estas son de utilidad para el diseño y forman un flujo natural y escalable para el desarrollo del algoritmo de determinación de correspondencias.

Los elementos más básicos en la estructura son los *segmentos* los cuales consisten en un par de puntos en la imagen, $\mathbf{p} = (p_x, p_y)$ y $\mathbf{q} = (q_x, q_y)$. Estos *segmentos* forman lo que son los lados del *cuadrilátero*, el próximo elemento estructural del marcador.

Un *cuadrilátero* o *quadrilateral* en inglés, al que se le denomina Ql , está determinado por cuatro segmentos conexos y distintos entre sí. El cuadrilátero tiene dos propiedades notables; el *centro* definido como el punto medio entre sus cuatro vértices y el *perímetro* definido como la suma de el largo de sus cuatro lados. Los *vértices* de un cuadrilátero se determinan mediante la intersección, en sentido amplio, de dos segmentos contiguos. Es decir, si s_1 es contiguo a s_2 dadas las recta r_1 que pasa por los puntos $(\mathbf{p}_1, \mathbf{q}_1)$ del segmento s_1 y la recta r_2 que pasa por los puntos $(\mathbf{p}_2, \mathbf{q}_2)$ del segmento s_2 , se determina el vértice correspondiente como la intersección $r_1 \cap r_2$.

A un *conjunto de cuadriláteros* o *quadrilateral set* se le denomina $QlSet$ y se construye a partir de M cuadriláteros, con $M > 1$. Los cuadriláteros comparten un mismo centro pero se diferencian en un factor de escala. A partir de dichos cuadriláteros se construye un lista ordenada $(Ql[0], Ql[1], \dots, Ql[M - 1])$ en donde el orden viene dado por el valor de perímetro de cada Ql . Se define el *centro del grupo de cuadriláteros*, \mathbf{c}_i , como el promedio de los centros de cada Ql de la lista ordenada.

Finalmente el *marcador QR* está constituido por N conjuntos de cuadriláteros dispuestos en una geometría particular. Esta geometría permite la determinación de un sistema de coordenadas; un origen y dos ejes a utilizar. Se tiene una lista ordenada $(QlSet[0], QlSet[1], \dots, QlSet[N - 1])$ en donde el orden se puede determinar mediante la disposición espacial de los mismos o a partir de hipótesis razonables.

Un marcador proveerá un numero de $4 \times M \times N$ vértices y por lo tanto la misma cantidad de puntos para proveer las correspondencias $\mathbf{M}_i \leftrightarrow \mathbf{m}_i$ al algoritmo de estimación de pose. De esta forma se tienen una cantidad de puntos superior a los que se podrían tener utilizando uno de los marcadores de los sistemas como *ARToolKit* a un costo de detección relativamente bajo. Por otro lado se podría agregar algún patrón para la identificación de marcadores en la esquina que completa el rectángulo en donde no hay $QlSet$ como se realizó en el trabajo Autoposicionamiento 3D [?].

4.2.2. Diseño

En base a las estructuras previamente definidas es que se describe el diseño del marcador. Como ya se explicó se toma un marcador tipo QR basado en cuadriláteros y más específicamente en tres conjuntos de tres cuadrados dispuestos en como se muestra en la figura 4.2.

Los tres cuadriláteros correspondientes a un mismo conjunto de cuadriláteros tienen idéntica alineación e idéntico centro. Los diferencia un factor de escala, esto es, $Ql[0]$ tiene lado l mientras que $Ql[1]$ y $Ql[2]$ tienen lado $2l$ y $3l$ respectivamente. Esto se puede ver en la figura 4.3. Adicionalmente se define un sistema de coordenadas con centro en el centro del $QlSet$ y ejes definidos como x horizontal a la derecha e y vertical hacia abajo. Esta convención en las direcciones de los ejes es

muy utilizada en el área de Procesamiento de Imágenes para definir las direcciones de los ejes de una imagen. Definido el sistema de coordenadas de puede fijar un orden a los vértices v_{j_1} de cada cuadrilátero $Ql[j]$ como,

$$\begin{aligned} v_{j_0} &= (a/2, a/2) & v_{j_2} &= (-a/2, -a/2) \\ v_{j_1} &= (a/2, -a/2) & v_{j_3} &= (-a/2, a/2) \end{aligned}$$

con $a = (j + 1) \times l$. El orden aquí explicado se puede ver también junto con el sistema de coordenadas en la figura 4.4.

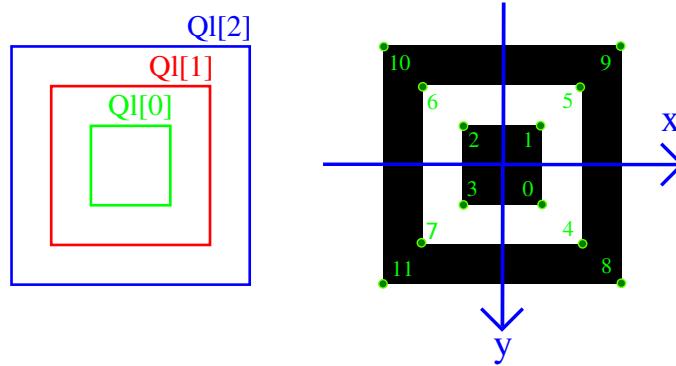


Figura 4.3: Detalle de un $QlSet$. A la izquierda se muestra el resultado de la detección de un $QlSet$ y el orden interno de sus cuadriláteros y a la derecha el orden de los vértices respecto al sistema de coordenadas local.

Un detalle del marcador completo se muestra en la figura 4.4 en donde se define el conjunto i de cuadriláteros concéntricos como el $QlSet[i]$ y se definen los respectivos centros de cada uno de ellos como \mathbf{c}_i . El sistema de coordenadas del marcador QR tiene centro en el centro del $QlSet[0]$ y ejes de coordenadas idénticos al definido para cada Ql . Se tiene además que los ejes de coordenadas pueden ser obtenidos mediante los vectores normalizados,

$$\mathbf{x} = \frac{\mathbf{c}_1 - \mathbf{c}_0}{\|\mathbf{c}_1 - \mathbf{c}_0\|} \quad \mathbf{y} = \frac{\mathbf{c}_2 - \mathbf{c}_0}{\|\mathbf{c}_2 - \mathbf{c}_0\|} \quad (4.1)$$

La disposición de los $QlSet$ es tal que la distancia indicada d_{01} definida como la norma del vector entre los centros \mathbf{c}_1 y \mathbf{c}_0 es significativamente mayor que la distancia d_{02} definida como la norma del vector entre los centros \mathbf{c}_2 y \mathbf{c}_1 . Esto es, $d_{01} \gg d_{02}$. Este criterio facilita la identificación de los $QlSet$ entre sí basados únicamente en la posición de sus centros y es explicado en la sección de determinación de correspondencias (sec.: 4.3.3).

4.2.3. Parámetros de diseño

Provisto el diseño del marcador descrito, quedan definidos ciertos parámetros **estructurales** que fueron de tomados fijos a lo largo del proyecto pero que podrían ser cambiados para trabajos futuros asociados. Estos parámetros son:

- M: cantidad de conjuntos de cuadriláteros.
- N: cantidad de cuadriláteros por conjuntos de cuadriláteros.
- Geometría: geometría de los cuadriláteros (Ql).

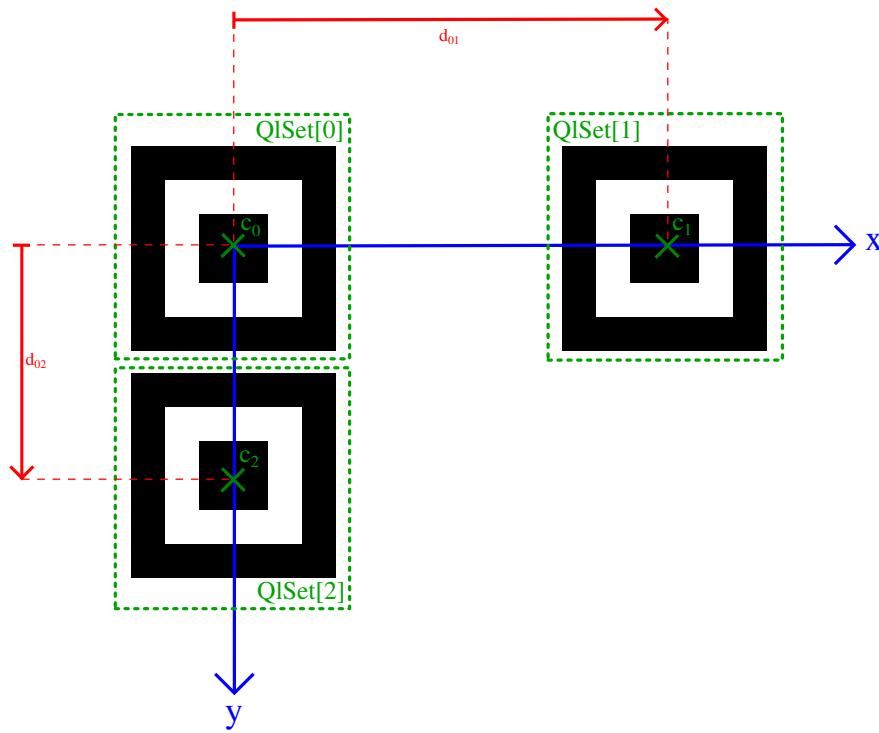


Figura 4.4: Detalle del marcador propuesto formando un sistema de coordenadas.

- Disposición: disposición espacial de los conjuntos de cuadriláteros ($QlSet$).

El criterio de elección de M y N parte del diseño los códigos QR como ya fue explicado. La detección por segmentos de línea resulta una cantidad de $3 \times QlSet$'s conteniendo $3 \times Ql$'s cada uno. Bajo esta elección de parámetros se tienen 36 segmentos y vértices. Se tiene entonces un número de puntos característicos razonable para la estimación de pose.

La elección de *cuadrados* como parámetro de geometría se basa en la necesidad de tener igual resolución en los dos ejes del marcador. De esta forma se asegura una distancia límite en donde, en un caso ideal enfrentado al marcador, la detección de segmentos de línea falla simultáneamente en los segmentos verticales como en los horizontales. De otra forma se tendría una dirección que limita más que la otra desaprovechando resolución.

La disposición espacial de los conjuntos de cuadriláteros está en primer lugar limitada a un plano y en segundo lugar es tal que se puede definir ejes de coordenadas ortogonales mediante los centros como se muestra en la figura 4.4.

Por otro lado se tiene otro juego de parámetros **dinámicos** que concluyen con el diseño del marcador. Estos parámetros conservan la estructura intrínseca del marcador permitiendo versatilidad en la aplicación y sin la necesidad de modificación alguna de los algoritmos desarrollados. Estos son:

- d_{ij} : distancia entre los centros $QlSet[j]$ con $QlSet[i]$.
- l : lado del cuadrilátero más pequeño ($Ql[0]$) de los $QlSet$.

En este caso se debe cumplir siempre la condición impuesta previamente en donde $d_{01} \gg d_{02}$. De otra forma se deberán realizar ciertas hipótesis no genéricas o se deberá aumentar ligeramente la complejidad del algoritmo para la identificación del marcador.

4.2.4. Diseños utilizados

- **Test:** Durante el desarrollo de los algoritmos de detección e identificación de los vértices del marcador QR se trabajó con determinados parámetros de diseño de dimensiones apropiadas para posibilitar el traslado y las pruebas domésticas.

- $l = 30\text{mm}$
- $d_{01} = 190\text{mm}$
- $d_{02} = 100\text{mm}$

- **Da Vinci**

- **Artigas**

- **Mapa**

4.3. Detección

La etapa de detección del marcador se puede separar en tres grandes bloques; la detección de segmentos de línea, el filtrado de segmentos y la determinación de correspondencias (figura ??). En esta sección se muestran algunos resultados para la detección de segmentos de línea por LSD y se desarrolla en profundidad los algoritmos desarrollados durante el proyecto para el filtrado de segmentos y determinación de correspondencias.

4.3.1. Detección de segmentos de línea

La detección de segmentos de línea se realiza mediante el uso del algoritmo LSD el cual se detalla en el capítulo ???. En forma resumida, dicho algoritmo toma como entrada una imagen en escala de grises de tamaño $W \times H$ y devuelve una lista de segmentos en forma de pares de puntos de origen y destino.

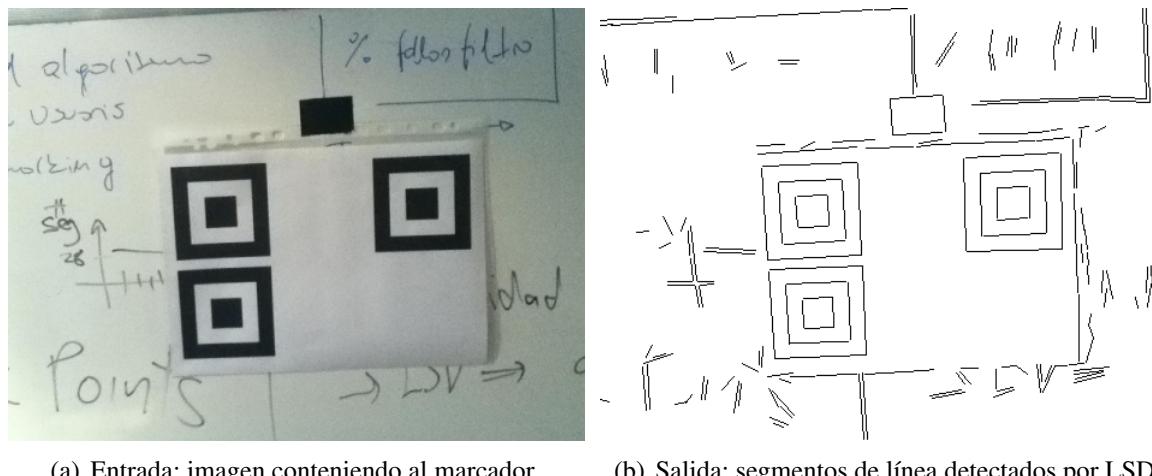


Figura 4.5: Resultados del algoritmo de detección de segmentos de línea LSD.

4.3.2. Filtrado y agrupamiento de segmentos

El filtrado y agrupamiento de segmentos consiste en la búsqueda de conjuntos de cuatro segmentos conexos en la lista de segmentos de línea detectados por LSD. Los conjuntos de segmentos conexos encontrados se devuelven en una lista en el mismo formato a la de LSD pero agrupados de a cuatro. A continuación se realiza una breve descripción del algoritmo de filtrado de segmentos implementado.

Se parte de una lista de m segmentos de línea,

$$\mathbf{L} = (\mathbf{s}_0 \ \mathbf{s}_1 \ \dots \ \mathbf{s}_{m-1})^t \quad (4.2)$$

y se recorre en i en busca de segmentos vecinos. La estrategia utilizada consiste en buscar, para el i -ésimo segmento \mathbf{s}_i , dos segmentos vecinos. En una primera etapa \mathbf{s}_j y en una segunda etapa \mathbf{s}_k , de forma que se forme una “U” como se muestra en la figura 4.6. La tercera etapa de búsqueda consiste en completar ese conjunto con un cuarto segmento \mathbf{s}_l que cierre la “U”.

Dos segmentos \mathbf{s}_i y \mathbf{s}_j son vecinos si se cumple que la distancia euclídea entre puntos, d_{ij} , es menor a un cierto umbral para alguna de las combinaciones $\mathbf{p}_i \leftrightarrow \mathbf{p}_j$, $\mathbf{q}_i \leftrightarrow \mathbf{q}_j$, $\mathbf{p}_i \leftrightarrow \mathbf{q}_j$ o $\mathbf{q}_i \leftrightarrow \mathbf{p}_j$. En la primera etapa de la búsqueda se testeán todas las posibilidades mientras que en la segunda etapa se testeán solo los puntos del segmento que no fueron utilizados. Por ejemplo, si se encontró la correspondencia $\mathbf{p}_i \leftrightarrow \mathbf{p}_j$ se busca el k -ésimo segmento \mathbf{s}_k que cumple que la distancia euclidiana d_{ik} es menor a cierto umbral para alguna de las combinaciones $\mathbf{q}_i \leftrightarrow \mathbf{p}_k$ y $\mathbf{q}_i \leftrightarrow \mathbf{q}_k$. En la tercera etapa la chequeo se realiza de forma aún más restringida probando para el segmento \mathbf{s}_l correspondencia simultánea entre sus puntos y solamente un punto cada uno de los segmentos \mathbf{s}_j y \mathbf{s}_k .

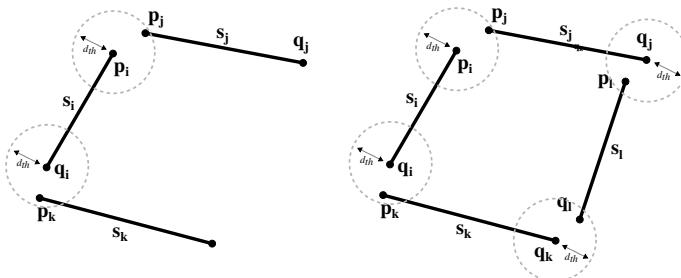
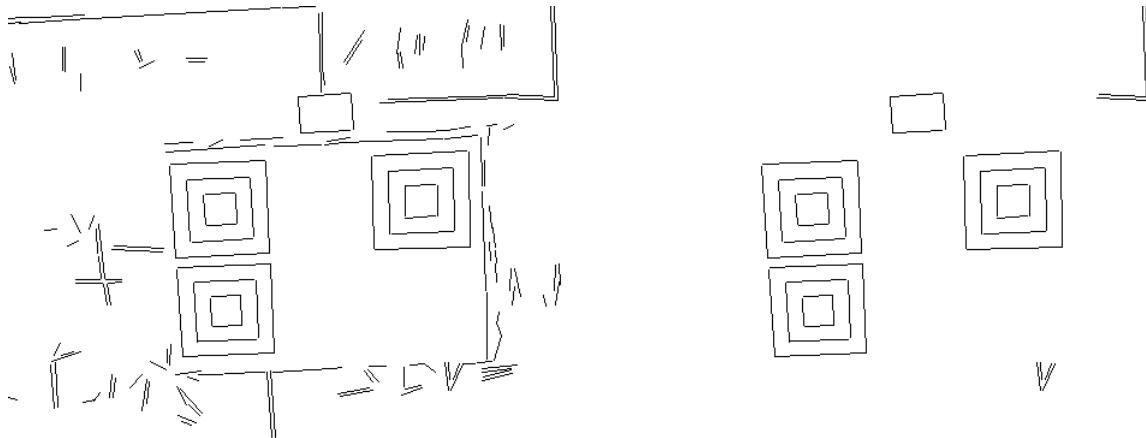


Figura 4.6: Conjunto de cuadriláteros conexos. A la izquierda la primera y segunda etapa del filtrado completadas para el segmento \mathbf{s}_i en donde se busca una “U”. A la derecha la última etapa en donde se cierra la “U” con el segmento \mathbf{s}_l .

Una vez encontrado el conjunto de cuatro segmentos conexos estos se marcan como utilizados, se guardan en una lista de salida y se continúa con el segmento $i + 1$ hasta recorrer los m segmentos de la lista de entrada. De esta forma se obtiene una lista de salida \mathbf{S} de n segmentos en donde n es por construcción múltiplo de cuatro.

En la figura 4.3.2 se muestran los resultados obtenidos para el algoritmo tomando como entrada la lista de segmentos de LSD. Se puede ver que los lados de los cuadrados del marcador son detectados correctamente pero también hay otras detecciones presentes. Por ejemplo el rectángulo negro correspondiente a un trozo de cinta negra que sostenía el marcador (ver figura ??(a)). También sobreviven otro tipo de elementos indeseados que se explican a continuación.

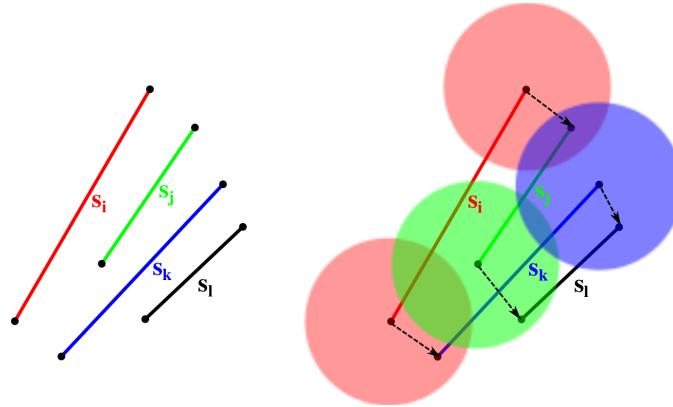
El algoritmo descrito es simple y provee resultados aceptables en general pero es propenso a tanto a detectar *falsos positivos* como al *sobre-filtrado* algunos conjuntos.



(a) Entrada: segmentos de línea detectados por LSD (b) Salida: segmentos de línea filtrados y agrupados

Figura 4.7: Resultados del algoritmo de filtrado y agrupamiento de segmentos de línea.

La detección de falsos positivos se puede atribuir principalmente a la condición de vecindad utilizada en donde un caso como el que se muestra en la figura 4.8 de un conjunto de segmentos paralelos cercanos y de tamaño similar “sobrevive” al filtrado de segmentos. De forma de evitar estos falsos positivos, se podría considerar implementar una condición de vecindad que tome en cuenta el punto de intersección entre los segmentos y la distancia de este punto a los puntos p , q más cercanos de cada segmento. Como se explicará en la sección ??, debido a que el algoritmo de determinación de correspondencias realiza la intersección entre estos segmentos se puede checar alguna condición sobre los segmentos o su intersección y en ese momento filtrar estos casos.

Figura 4.8: Posible configuración de segmentos paralelos que “sobreviven” al filtrado. A la izquierda el grupo de segmentos, a la derecha se muestra como se desarrolla el filtrado de s_i .

El sobre-filtrado de segmentos tiende a ocurrir cuando no se cumple la condición de distancia entre segmentos vecinos cuando visualmente si lo son. Se debe principalmente a que se utiliza para el filtrado un valor de d_{th} fijo que resulta en buenos resultados para la aplicación pero en ciertas circunstancias produce este problema. Esta medida de distancia se podría tomar relativa al largo del los segmentos a *testear* de forma de generalizar el valor pero se debería analizar un poco más en detalle la posible implementación para que resulte en buenos resultados y no introduzca otra clase de errores.

El algoritmo de filtrado y agrupamiento de segmentos es sensible respecto a la elección del parámetro d_{th} . Si esta parámetro está por debajo del valor óptimo la *performance* del algoritmo

se verá afectada fuertemente pues se corre el riesgo de sobre filtrar y no proporcionar suficientes segmentos para la correcta determinación de correspondencias. Por el contrario, si el parámetro está por encima del valor óptimo, el filtrado tiende a proveer falsos positivos aun que este caso no llega a ser tan crítico como el primero para la aplicación. A modo de ejemplo, para una imagen de tamaño 480×320 con el marcador ocupando entre un 25 % y un 80 % el valor del parámetro que da mejores resultados es aproximadamente de 6 o 7 píxeles.

4.3.3. Determinación de correspondencias

Se detalla a continuación el algoritmo de determinación de correspondencias a partir de grupos de cuatro segmentos de línea conexos. Para ese algoritmo se hace uso de los elementos estructurales del marcado (sec.: 4.2.1), de forma de desarrollar un algoritmo modular, escalable y simple.

Se toma como entrada la lista de segmentos filtrados y agrupados

$$\mathbf{S} = (\mathbf{s}_0 \ \mathbf{s}_1 \ \dots \ \mathbf{s}_i \ \mathbf{s}_{i+1} \ \mathbf{s}_{i+2} \ \mathbf{s}_{i+3} \ \dots \ \mathbf{s}_{n-1})^t \quad (4.3)$$

en donde cada segmento se compone de un punto inicial \mathbf{p}_i y un punto final \mathbf{q}_i , $\mathbf{s}_i = (\mathbf{p}_i, \mathbf{q}_i)$, con n múltiplo de cuatro. Si i también lo es, entonces el sub-conjunto, $\mathbf{S}_i = (\mathbf{s}_i \ \mathbf{s}_{i+1} \ \mathbf{s}_{i+2} \ \mathbf{s}_{i+3})^t$, corresponde a un conjunto de cuatro segmentos del línea conexos.

Para cada sub-conjunto \mathbf{S}_i se intersectan entre sí los segmentos obteniendo una lista de cuatro vértices, $\mathbf{V}_i = (\mathbf{v}_i \ \mathbf{v}_{i+1} \ \mathbf{v}_{i+2} \ \mathbf{v}_{i+3})^t$. Si \mathbf{r}_i es la recta que pasa por los puntos \mathbf{p}_i y \mathbf{q}_i del segmento \mathbf{s}_i , la lista de vértices se obtiene como sigue,

$$\begin{aligned} \mathbf{v}_i &= \mathbf{r}_i \cap \mathbf{r}_{i+1} \\ \mathbf{v}_{i+1} &= \mathbf{r}_i \cap \mathbf{r}_{i+2} \\ \mathbf{v}_{i+2} &= \mathbf{r}_{i+3} \cap \mathbf{r}_{i+2} \\ \mathbf{v}_{i+3} &= \mathbf{r}_{i+3} \cap \mathbf{r}_{i+1} \end{aligned}$$

resultando en dos posibles configuraciones de vértices. Las dos configuraciones se muestran en la figura 4.9 en donde una de ellas tiene sentido horario y la otra antihorario partiendo de v_i .

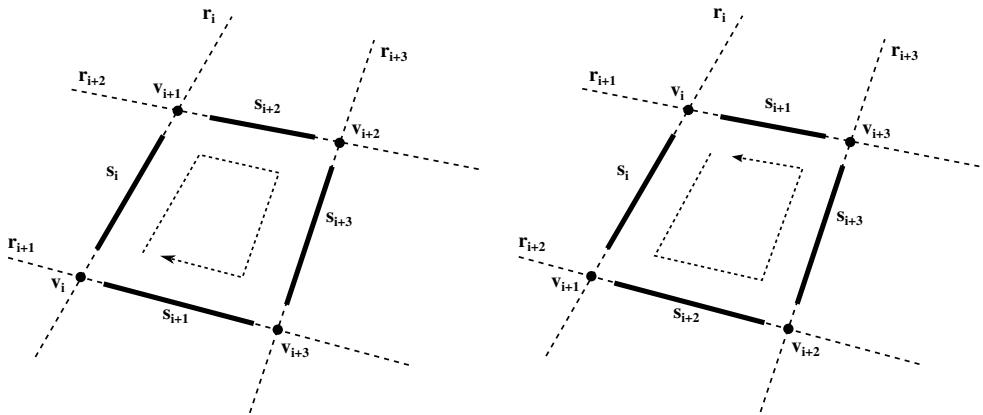


Figura 4.9: Posibles configuraciones de vértices posterior a la intersección de conjuntos de segmentos pertenecientes a un cuadrilátero.

Posterior a la intersección se realiza un chequeo sobre el valor de las coordenadas de los vértices. Si alguno de ellos se encuentra fuera de los límites de la imagen, el conjunto de cuatro segmentos es marcado como inválido. Este chequeo resulta en el filtrado de “falsos cuadriláteros” corrigiendo

un defecto del filtrado de segmentos, como por ejemplo un grupo de segmentos paralelos cercanos como ya se explicó.

Para cada uno de los conjuntos de vértices se construye con ellos un elemento cuadrilátero que se almacena en una lista de cuadriláteros

$$QlList = (Ql[0] \quad Ql[1] \quad \dots \quad Ql[i] \quad \dots \quad Ql[\frac{n}{4}])^t$$

A partir de esa lista de cuadriláteros, se buscan grupos de tres cuadriláteros $QlSet$ que “compartan” un mismo centro. Para esto se recorre ordenadamente la lista en i buscando para cada cuadrilátero dos cuadriláteros j y k que cumplan que la distancia entre sus centros y el del i -ésimo cuadrilátero sea menor a cierto umbral c_{th} ,

$$d_{ij} = \|\mathbf{c}_i - \mathbf{c}_j\| < c_{th}, \quad d_{ik} = \|\mathbf{c}_i - \mathbf{c}_k\| < c_{th}. \quad (4.4)$$

Estos cuadriláteros se marcan en la lista como utilizados con ellos se forma el l -ésimo $QlSet$ ordenándolos según su perímetro, de menor a mayor como

$$QlSet[l] = (Ql[0] \quad Ql[1] \quad Ql[2])$$

con $l = (0, 1, 2)$. Esta búsqueda se realiza hasta encontrar un total de tres $QlSet$ completos de forma de obtener un marcador completo, esto es, detectando todos los cuadriláteros que lo componen.

Una vez obtenida la lista de tres $QlSet$,

$$QlSetList = (QlSet[0] \quad QlSet[1] \quad QlSet[2])$$

ésta se ordena de forma que su disposición espacial se corresponda con la del marcador QR. Para esto se calculan las distancias entre los centros de cada $QlSet$ y se toma el índice i como el índice que produce el vector de menor distancia, $\mathbf{u}_i = \mathbf{c}_{i+1} - \mathbf{c}_i$. En este punto que es importante que la condición de distancia entre los centros de los $QlSet$ se cumpla, $d_{10} \gg d_{20}$, para una simple identificación. Bajo una transformación proyectiva del marcador, es posible que esta relación se modifique e incluso que deje de valer pero imponiendo la condición “mucho mayor” se asegura que el algoritmo funciona correctamente para condiciones razonables. Esto es, para proyecciones o poses que se encuentran dentro de las hipótesis uso de la aplicación.

Una vez seleccionado el vector \mathbf{u}_i , se tienen obtiene el juego de vectores $(\mathbf{u}_i, \mathbf{u}_{i+1}, \mathbf{u}_{i+2})$ como se muestra en la figura 4.10.

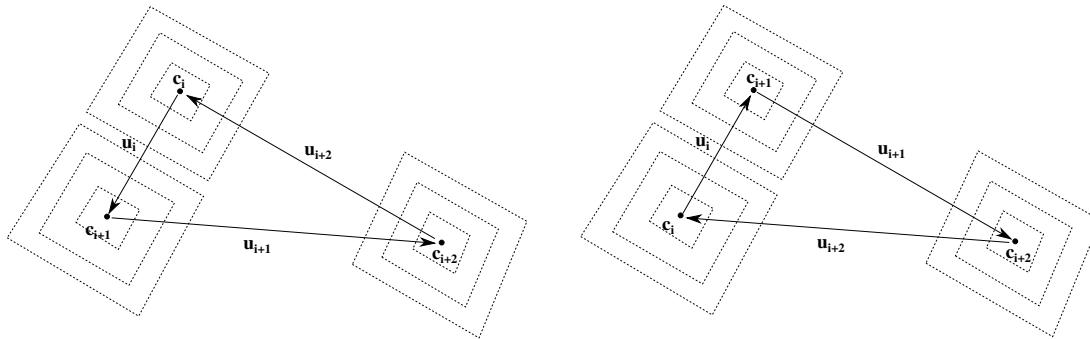


Figura 4.10: Vértices de cada Ql ordenados respecto al signo de sus proyecciones contra el sistema de coordenadas local a cada $QlSet$.

Existen solo dos posibles configuraciones para estos vectores por lo que se utiliza este conocimiento para ordenar los $QlSet$ de la lista realizando el producto vectorial, aumentando la dimensión de los vectores $\hat{\mathbf{u}}_i$ y $\hat{\mathbf{u}}_{i+1}$ con coordenada $z = 0$,

$$\mathbf{b} = \hat{\mathbf{u}}_i \times \hat{\mathbf{u}}_{i+1}.$$

Si el vector \mathbf{b} tiene valor en la coordenada z positivo se ordena como,

$$\begin{aligned} QlSet[0] &\leftarrow QlSet[i] \\ QlSet[1] &\leftarrow QlSet[i+2] \\ QlSet[2] &\leftarrow QlSet[i+1] \end{aligned}$$

o de lo contrario se ordena como,

$$\begin{aligned} QlSet[0] &\leftarrow QlSet[i+1] \\ QlSet[1] &\leftarrow QlSet[i+2] \\ QlSet[2] &\leftarrow QlSet[i] \end{aligned}$$

Por ultimo se construye un marcador QR que contiene la lista de tres $QlSet$ ordenados según lo indicado permitiendo la definición de un centro de coordenadas como el centro \mathbf{c}_0 del $QlSet[0]$ y ejes de coordenadas definidos en la ecuación 4.1. Los ejes de este sistema de coordenadas permiten, para cada Ql de cada $QlSet$, proyectar los vértices sobre el sistema de coordenadas local al $QlSet$ y según su signo ordenarlos como se muestra en la figura 4.11. De esta forma, recorriendo ordenada-

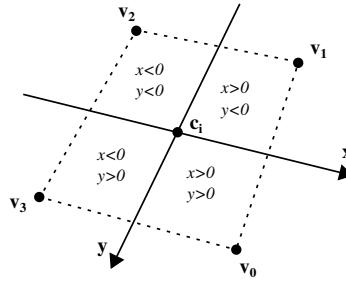


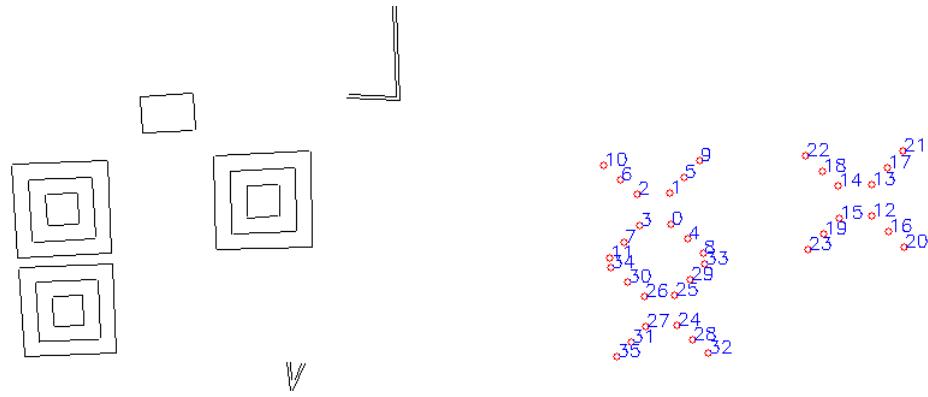
Figura 4.11: Posibles configuraciones de centros resultan en la orientación de los vectores \mathbf{u}_{i+k} .
mente los elementos del marcador, se ordenan los vértices de cada Ql del marcador.

Por último, a partir del marcador ordenado, se extrae una lista de vértices que se corresponde con la lista de vértices del marcador en coordenadas del mundo. Este recorrido se realiza en el siguiente orden,

```

for  $i = (0, 1, 2)$  do
  for  $j = (0, 1, 2)$  do
    for  $k = (0, 1, 2, 3)$  do
      So obtiene el punto vértice:  $\mathbf{p} = QlSet[i] \rightarrow Ql[j] \rightarrow v[k];$ 
      Se agrega a la lista de correspondencias  $\mathbf{m}_l \leftarrow \mathbf{p};$ 
      Se incrementa  $l;$ 
  
```

Se determinan las correspondencias $\mathbf{M}_i \leftrightarrow \mathbf{m}_i$ necesarias para la estimación de pose las cuales se muestran en la figura 4.3.3. Se puede ver que el algoritmo de determinación de correspondencias funciona correctamente por lo que los “falsos” cuadriláteros que sobreviven al filtrado de segmentos no son un problema.



(a) Entrada: segmentos de línea filtrados y agrupados (b) Salida: puntos vértices ordenados.

Figura 4.12: Resultados del algoritmo de determinación de correspondencias.

4.3.4. Detección robusta

El algoritmo descripto al momento requiere que dentro de la lista de segmentos filtrados se encuentren todos los segmentos que componen el marcador pero este requerimiento representa un problema importante en cuanto a el desempeño del algoritmo. En caso de que esto no se cumpla no es posible proporcionar las correspondencias necesarias para la estimación de pose y no se tendrá una pose válida para ese cuadro o *frame* para la aplicación. En aplicaciones en tiempo real en donde el procesamiento de la imagen es la mayor limitante, la fluidez visual dada por el *frame rate* se ve notablemente perjudicada resultando en que el sistema sea incómodo e incluso inutilizable. Es por esto que en esta sección se desarrolla la extensión del algoritmo de determinación de correspondencias para una cantidad menor de segmentos detectados y filtrados que resulta en una mejor sustancial en la cantidad de *frames* en los cuales es posible determinar correspondencias y obtener así una pose válida.

Se busca una determinación de correspondencias más robusta pero manteniendo las esencia del algoritmo desarrollado. Por esto se tienen dos aspectos a tomar en cuenta; la detección de *QlSet*'s se realiza basada en la búsqueda de cuadriláteros concéntricos por lo que se debe contar con un mínimo de dos cuadriláteros por *QlSet* para permitir la diferenciación entre un conjunto de segmentos filtrados debido a que pertenecen al marcador y a otro conjunto que no pertenece pero si cumple con las condiciones, por ejemplo podría ser el marco de una obra o cualquier elemento en la escena que forme un cuadrilátero. Esto fija un límite de no menos de 24 segmentos necesarios para el funcionamiento. El otro aspecto a tomar en cuenta se refiere a la forma en que se ordenan los *Ql*'s dentro de cada *QlSet*. Como ya se explicó el orden se basa en la medida del perímetro de los *Ql*'s ordenando de menor a mayor por lo que será necesario contar con, al menos, un *QlSet* completo de forma de tener una referencia a la hora de identificar los *QlSet*'s incompletos hallados. Por lo tanto la extensión del algoritmo permite una correcta identificación de los vértices del marcador con un número mayor o igual a 28 segmentos.

La implementación de esta extensión del algoritmo se realizó manteniendo la estructura básica descrita anteriormente y se detalla aquí solamente los agregados realizados.

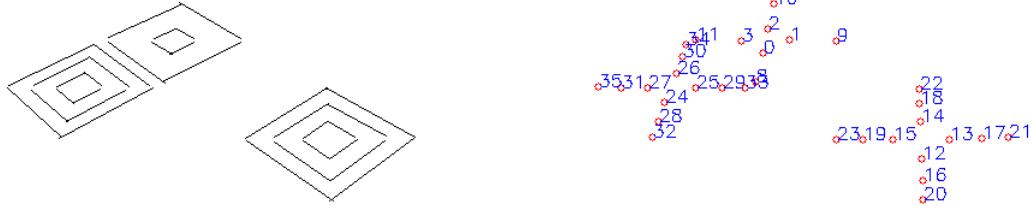
Al realizar la búsqueda de conjuntos de cuadriláteros concéntricos se buscan en primer lugar los *QlSet*'s completos y luego en caso de que estos no lleguen a ser tres, se intenta completar buscando *QlSet*'s incompletos o sea conjuntos de dos cuadriláteros que comparten un mismo centro. Estos se

agrupan en una lista de la misma forma en que se describió anteriormente pero dejando el tercer cuadrilátero, $Ql[2]$, marcado como inválido.

Una vez completada la lista de tres $QlSet$, con al menos uno de ellos detectado completo, se ordenan en primer lugar los $QlSet$ completos y de ellos se extrae una lista de perímetros promedio. Esta lista de perímetros promedio se utiliza para el ordenamiento de los $QlSet$ incompletos comparando con los perímetros de los $Ql[0]$ y $Ql[1]$ de cada $QlSet$. El $Ql[2]$ previamente marcado como inválido se posiciona por descarte en la posición que corresponda.

Al momento de proporcionar la lista de vértices ordenados \mathbf{m}_i y correspondientes con los del modelo \mathbf{M}_i , se introducen valores inválidos para los Ql 's marcados como inválidos. Por último se realiza un recorte de las dos listas de puntos en base a estos valores inválidos, se recorre la lista de puntos en la imagen \mathbf{m}_i y se extraen de la lista de puntos en la imagen y de los puntos del modelo los puntos inválidos obteniendo un juego de al menos 28 correspondencias $\mathbf{m}'_i \leftrightarrow \mathbf{M}'_i$ para el algoritmo de estimación de pose.

En la figura 4.3.4(a) se muestran imágenes en la que falla el filtrado de segmentos para uno de los cuadriláteros mientras que en la figura 4.3.4(b) se puede ver como el algoritmo de determinación de correspondencias provee 32 correspondencias ordenadas correctamente, diferenciando en el $QlSet$ incompleto los vértices.



(a) Entrada: segmentos de línea filtrados y agrupados

(b) Salida: puntos vértices ordenados.

Figura 4.13: Resultados del algoritmo de determinación de correspondencias robusto para una falla en el filtrado de segmentos.

4.3.5. Resultados

Mas imágenes con resultados?

Todo sobre la misma imagen de entrada?

O toda la secuencia para distintas imágenes?

CAPÍTULO 5

LSD: “Line Segment Detection”

5.1. Introducción

LSD es un algoritmo de detección de segmentos publicado recientemente [?]. Es temporalmente lineal, tiene precisión inferior a un píxel y no requiere de un ajuste previo de parámetros, como casi todos los demás algoritmos de idéntica función. Puede ser considerado el estado del arte en cuanto a detección de segmentos en imágenes digitales. Como cualquier otro algoritmo de detección de segmentos, LSD basa su estudio en la búsqueda de contornos angostos dentro de la imagen. Estos son regiones en donde el nivel de brillo de la imagen cambia notoriamente entre píxeles vecinos, lo cual puede ser detectado mediante el módulo del gradiente de la misma.

Se genera en primer lugar, un campo de orientaciones asociadas a cada uno de los píxeles denominado por los autores *level-line orientation field*. Dicho campo se obtiene de calcular las orientaciones ortogonales a los ángulos asociados al gradiente de la imagen. Luego, LSD puede verse como una composición de tres pasos:

- (1) División de la imagen en las llamadas *line-support regions*, que son grupos conexos de píxeles con idéntica orientación, a menos de cierta tolerancia.
- (2) Búsqueda del segmento que mejor aproxime cada *line-support region*: aproximación de las regiones por rectángulos.
- (3) Validación o no de cada segmento detectado en el punto anterior.

Los puntos (1) y (2) están basados en el algoritmo de detección de segmentos de Burns, Hanson y Riseman [?], y el punto (3) es una adaptación del método *a contrario* de Desolneux, Moisan y Morel [?].

En el presente capítulo se estudiará a fondo el algoritmo y se presentarán y justificarán algunos cambios que hubo que hacerle a la implementación del mismo con la que se contaba, versión 1.6 descargada de [?], para mejorar su desempeño en el tiempo real.

5.2. *Line-support regions*

El primer paso de LSD es el dividir la imagen en regiones conexas de píxeles con igual orientación, a menos de cierta tolerancia τ , llamadas *line-support regions*. El método para realizar tal división es del tipo “*region growing*”; cada región comienza por un píxel y cierto ángulo asociado, que en este caso coincide con el de este primer píxel. Luego, se testeán sus ocho vecinos y los que cuenten con un ángulo similar al de la región son incluídos en la misma. En cada iteración el ángulo asociado a la región es calculado como el promedio de las orientaciones de cada píxel dentro de la *line-support region*; la iteración termina cuando ya no se pueden agregar más píxeles a la misma.

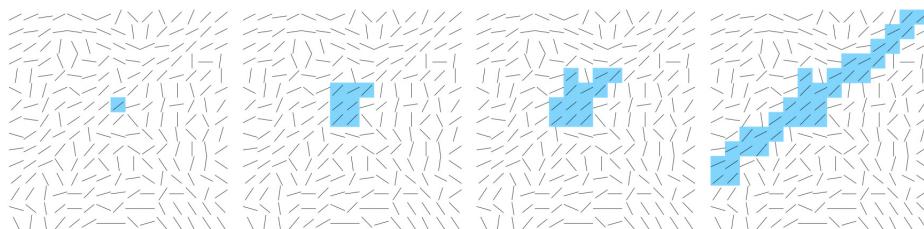


Figura 5.1: Proceso de crecimiento de una región. El ángulo asociado cada píxel de la imagen está representado por los pequeños segmentos y los píxeles coloreados representan la formación de la región. Tomada de [?].

Los píxeles agregados a una región son marcados de manera que no vuelvan a ser testeados. Para mejorar el desempeño del algoritmo, las regiones comienzan a evaluarse por los píxeles con gradientes de mayor amplitud ya que estos representan mejor los bordes.

Existen algunos casos puntuales en los que el proceso de búsqueda de *line-support regions* puede arrojar errores. Por ejemplo, cuando se tienen dos segmentos que se juntan y que son colineales a no ser por la tolerancia τ descripta anteriormente, se detectarán ambos segmentos como uno solo; ver Figura 5.2. Este potencial problema es heredado del algoritmo de Burns, Hanson y Riseman.



Figura 5.2: Potencial problema heredado del algoritmo de Burns, Hanson y Riseman. Izq.: Imagen original. Ctro.: Segmento detectado. Der.: Segmentos que deberían haberse detectado. Tomada de [?].

Sin embargo, LSD plantea un método para solucionar este tipo de problemas. Durante el proceso de crecimiento de las regiones, también se realiza la aproximación rectangular a dicha región (paso (2) de los tres definidos anteriormente); y si menos de cierto porcentaje umbral de los píxeles dentro del rectángulo corresponden a la *line-support region*, lo que se tiene no es un segmento. Se detiene entonces el crecimiento de la región.

5.3. Aproximación de las regiones por rectángulos

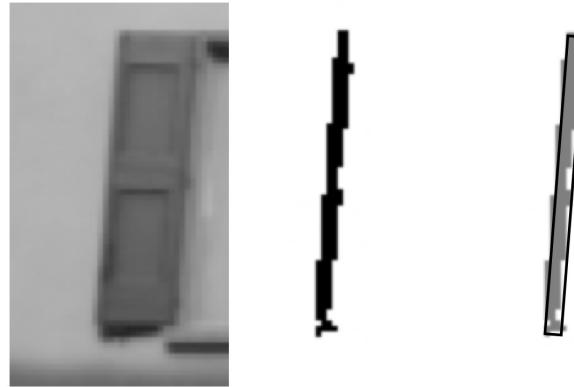


Figura 5.3: Búsqueda del segmento que mejor aproxime cada *line-support region*: aproximación de una región por un rectángulo. Izq.: Imagen original. Ctro.: Una de las regiones computadas. Der.: Aproximación rectangular que cubre el 99 % de la masa de la región. Tomada de [?].

Cada *line-support region* debe ser asociada a un segmento. Cada segmento será determinado por su centro, su dirección, su anchura y su longitud. A diferencia de lo que pudiese resultar intuitivo, la dirección asociada al segmento no se corresponde con la asociada a la región (el promedio de las direcciones de cada uno de los píxeles). Sin embargo, se elige el centro del segmento como el centro de masa de la región y su dirección como el eje de inercia principal de la misma; la magnitud del gradiente asociado a cada píxel hace las veces de masa. La idea detrás de este método es que los píxeles con un gradiente mayor en módulo, tienen una mayor probabilidad de corresponder a un borde. La anchura y la longitud del segmento son elegidos de manera de cubrir el 99 % de la masa de la región.

5.4. Validación de segmentos

La validación de los segmentos previamente detectados se plantea como un método de test de hipótesis. Se utiliza un modelo *a contrario*. El término *a contrario* viene del latín y significa “al revés” o “de forma opuesta”. En procesamiento de imágenes, el principio para la detección *a contrario* define, en primer lugar, un modelo llamado “*a priori*” para el caso genérico en el que no haya nada que detectar. Entonces la detección de un evento en particular sólo se dará cuando la cantidad de ocurrencias de dicho evento en el modelo *a priori* sea lo suficientemente baja. Nótese la aparición de cierto valor umbral a ajustar.

Para el caso de LSD, dada una imagen de ruido blanco y Gaussiano, se sabe que cualquier tipo de estructura detectada sobre la misma será casual. En rigor, se sabe que para cualquier imagen de este tipo, su *level-line orientation field* toma, para cada píxel, valores independientes y uniformemente distribuidos entre $[0, 2\pi]$. Dado entonces un segmento en la imagen analizada, se estudia la probabilidad de que dicha detección se dé en la imagen de ruido, y si ésta es lo suficientemente baja, el segmento se considerará válido, de lo contrario se considerará que se está bajo la hipótesis H_0 : un conjunto aleatorio de píxeles que casualmente se alinearon de manera de detectar un segmento.

Para estudiar la probabilidad de ocurrencia de una cierta detección en la imagen de ruido, se deben tomar en cuenta todos los rectángulos potenciales dentro de la misma. Dada una imagen

$N \times N$, habrán N^4 orientaciones posibles para los segmentos, N^2 puntos de inicio y N^2 puntos de fin. Si se consideran N posibles valores para la anchura de los rectángulos, se obtienen N^5 posibles segmentos. Por su parte, dado cierto rectángulo r , detectado en la imagen x , se denota $k(r,x)$ a la cantidad de píxeles alineados dentro del mismo. Se define además un valor llamado *Number of False Alarms* (NFA) que está fuertemente relacionado con la probabilidad de detectar al rectángulo en cuestión en la imagen de ruido X :

$$NFA(r,x) = N^5 \cdot P_{H_0}[k(r,X) \geq k(r,x)]$$

véase que el valor se logra al multiplicar la probabilidad de que un segmento de la imagen de ruido, de tamaño igual a r , tenga un número mayor o igual de píxeles alineados que éste, por la cantidad potencial de segmentos N^5 . Cuanto menor sea el número NFA, más significativo será el segmento detectado r ; pues tendrá una probabilidad de aparición menor en una imagen sin estructuras. De esta manera, se descartará H_0 , o lo que es lo mismo, se aceptará el segmento detectado como válido, si y sólo si:

$$NFA(r) \leq \varepsilon$$

donde empíricamente $\varepsilon = 1$ para todos los casos.

Si se toma en cuenta que cada píxel de la imagen ruidosa toma un valor independiente de los demás, se concluye que también lo harán su gradiente y su *level-line orientation field*. De esta manera, dada una orientación aleatoria cualquiera, la probabilidad de que uno de los píxeles de la imagen cuente con dicha orientación, a menos de la ya mencionada tolerancia τ , será:

$$p = \frac{\tau}{\pi}$$

además, se puede modelar la probabilidad de que cierto rectángulo en la imagen ruidosa, con cualquier orientación, formado por $n(r)$ píxeles, cuente con al menos $k(r)$ de ellos alineados, como una distribución binomial:

$$P_{H_0}[k(r,X) \geq k(r,x)] = B(n(r), k(r), p).$$

Finalmente, el valor *Number of False Alarms* será calculado para cada segmento detectado en la imagen analizada de la siguiente manera:

$$NFA(r,x) = N^5 \cdot B(n(r), k(r), p);$$

si dicho valor es menor o igual a $\varepsilon = 1$, el segmento se tomará como válido; de lo contrario se descartará.

5.5. Refinamiento de los candidatos

Por lo que se vió hasta el momento, la mejor aproximación rectangular a una *line-support region* es la que obtenga un valor NFA menor. Para los segmentos que no son validados, se prueban algunas variaciones a la aproximación original con el objetivo de disminuir su valor NFA y así entonces validarlos. Esta claro que este paso no es significativo para segmentos largos y bien definidos, ya que estos serán validados en la primera inspección; sin embargo, ayuda a detectar segmentos más pequeños y algo ruidosos.

Lo que se hace es probar distintos valores para la anchura del segmento y para sus posiciones laterales, ya que estas son los parámetros peor estimados en la aproximación rectangular, pero tienen un efecto muy grande a la hora de validar los segmentos. Es que un error de un píxel en el ancho de un segmento, puede agregar una gran cantidad de píxeles no alineados a este (tantos como el largo del segmento), y esto se ve reflejado en un valor mayor de NFA y puede llevar a una no detección.

Otro método para el refinamiento de los candidatos es la disminución de la tolerancia τ . Si los puntos dentro del rectángulo efectivamente corresponden a un segmento, aunque la tolerancia disminuya, se computará prácticamente misma cantidad de segmentos alineados; y con una probabilidad menor de ocurrencia ($\frac{\tau}{\pi}$), el valor NFA obtenido será menor. Los nuevos valores testeados de tolerancia son: $\frac{\tau}{2}$, $\frac{\tau}{4}$, $\frac{\tau}{8}$, $\frac{\tau}{16}$ y $\frac{\tau}{32}$. El nuevo valor NFA asociado al segmento será el menor de todos los calculados.

5.6. Optimización del algoritmo para tiempo real

Que un algoritmo de procesamiento de imágenes digitales sea temporalmente lineal significa que su tiempo de ejecución crece linealmente con el tamaño de la imagen en cuestión. Estos algoritmos son los mejores para el procesamiento de imágenes en tiempo real. Si bien, como se explicó con anterioridad, los autores de LSD afirman que este es temporalmente lineal; la implementación con la que se cuenta no fue pensada para ser ejecutada en tiempo real. Así entonces, para poder aumentar la tasa de cuadros por segundo total de la aplicación, hubo que realizar algunos cambios mínimos en el código, siempre buscando que estos alteren lo menos posible el desempeño del algoritmo. Se trabajó sobre ciertos bloques en particular.

5.6.1. Filtro Gaussiano

Antes de procesar la imagen con el algoritmo tal y como se vió en secciones anteriores, la misma es filtrada con un filtro Gaussiano. Se busca en primer lugar, disminuir el tamaño de la imagen de entrada con el objetivo de disminuir el volumen de información procesada. Además, al difuminar la imagen, se conservan únicamente los bordes más pronunciados. Para este proyecto en particular, se escogió la escala del submuestreo fija en 0,5, un poco más adelante en la corriente sección se explicará por qué.

Como la función Gaussiana 2D es separable, el filtrado de la imagen se hace en dos pasos, primero a lo ancho y luego a lo largo. Se utiliza el núcleo Gaussiano de una dimensión normalizado de la Figura 5.4.

De esta manera, se crea una imagen auxiliar vacía y escalada en x pero no en y , y se recorre asignándole a cada píxel en x su valor correspondiente, obtenido del promedio del píxel $\frac{x}{escala}$ en la imagen original y sus vecinos, todos ponderados por el núcleo Gaussiano centrado en $\frac{x}{escala}$. Luego se crea otra imagen, pero esta vez escalada tanto en x como en y , y se recorre asignándole a cada píxel en y su valor correspondiente, obtenido del promedio del píxel $\frac{y}{escala}$ en la imagen auxiliar y sus vecinos, todos ponderados por el núcleo Gaussiano centrado en $\frac{y}{escala}$. En la Figura 5.5 se muestra la relación entre las imágenes.

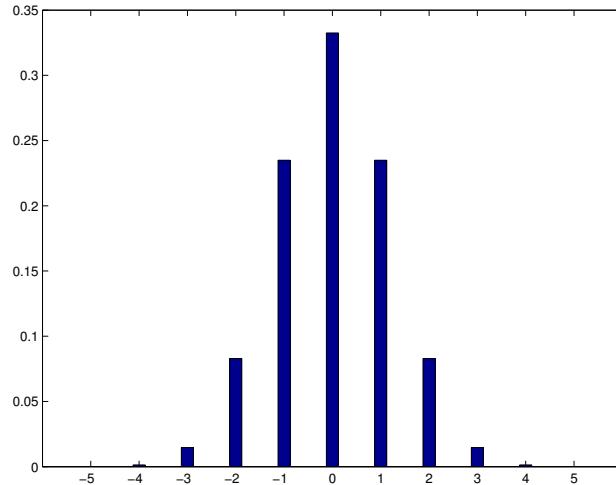


Figura 5.4: Núcleo Gaussiano utilizado por LSD. $\sigma = 1, 2$.

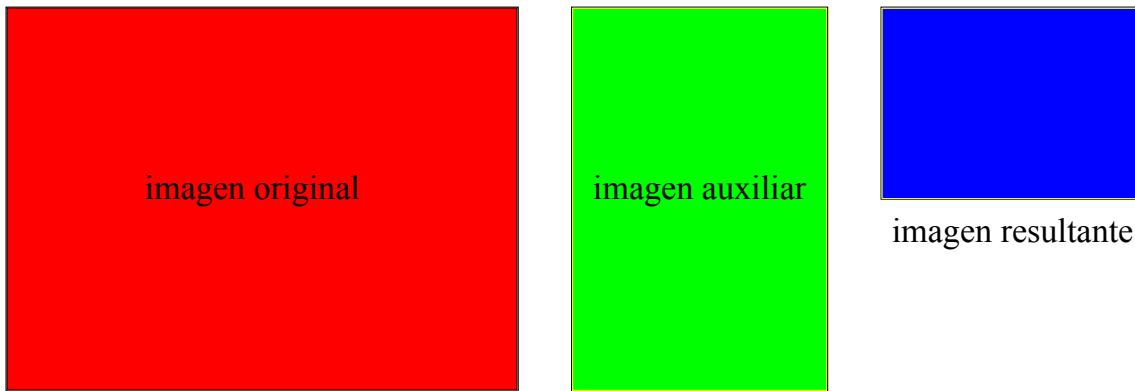


Figura 5.5: Relación entre las imágenes en consideradas en el filtro Gaussiano. Escala: 0,5.

Véase que cuando en el submuestreo $\frac{1}{escala}$ no es un entero, el centro del núcleo Gaussiano no siempre debe caer justo sobre un píxel en particular en la imagen original, sino que debe hacerlo entre dos de ellos. Lo que se hace entonces es mover $\pm 0,5$ píxeles al centro del núcleo en cada asignación de los píxeles en las imágenes escaladas; de manera de que la ponderación en el promediado de los píxeles de la imagen original (y luego la auxiliar) sea la debida. Aunque esta operación le agrega precisión al algoritmo, también le agrega un gran costo computacional, ya que lo que se hace es crear un nuevo núcleo Gaussiano en cada caso. En particular, para una imagen escalada de 240×180 píxeles (dimensiones efectivamente utilizadas en este proyecto), debido al filtrado en dos pasos, el núcleo Gaussiano se crea y se destruye $86400 + 43200 = 129600$ veces.

Se decidió redondear la escala de submuestreo en 0,5, ya que los valores utilizados empíricamente hasta el momento rondaban este valor, y se concluyó que para dicha escala, el núcleo Gaussiano debía permanecer constante, siempre centrado en su sexta muestra (ver Figura 5.4); por lo que se lo quitó de la iteración y actualmente se crea una sola vez al ingresar la imagen al filtro. Es importante destacar que esta optimización es transparente para el algoritmo si y sólo si $\frac{1}{escala} = n$, donde n es un entero.

Otro cambio que se le realizó al filtrado Gaussiano fue la supresión de las condiciones de borde.

Cuando se filtra cualquier imagen con un filtro con memoria, algo importante a tener en cuenta son las condiciones de borde, ya que para el procesamiento de los extremos de la imagen, estos filtros requieren de píxeles que están fuera de sus límites. Algunas de las soluciones a este problema son periodizar la imagen, simetrizarla o hasta asumir el valor 0 para los píxeles que estén fuera de esta. La opción escogida por LSD es la simetrización. Demás está decir que este proceso requiere de cierto costo computacional extra, por lo que se lo decidió suprimir. Este costo computacional extra se debe a que el algoritmo encargado del filtrado debe estar en cada bucle preguntándose si es necesario contar con el valor de algún píxel fuera de los límites de la imagen, y en ese caso asignarle a dicho píxel el valor de su correspondiente simétrico, con eje de simetría el borde de la imagen más próximo. Actualmente, la imagen escalada no es computada en sus píxeles terminales; estos son 3 al inicio de cada línea o columna y 2 al final de cada una de ellas, irrelevantes en el tamaño total de la imagen y también, por ser un filtro FIR (“Finite Impulse Response”), en el resultado del filtrado en general. Ver Figura 5.6.

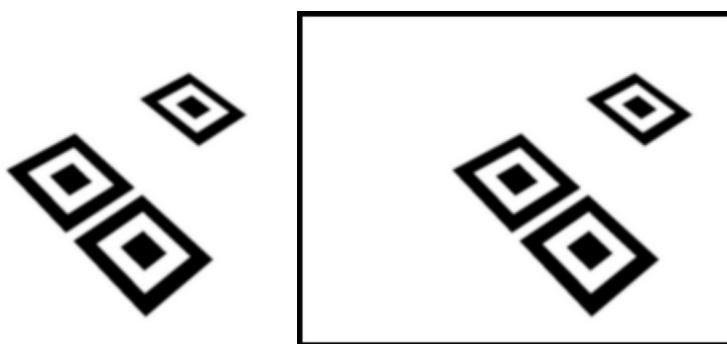


Figura 5.6: Imagen artificial del marcador trasladado y rotado, filtrada con el filtro Gaussiano. Izq.: Filtro Original. Der.: Filtro sin las condiciones de borde.

5.6.2. *Level-line angles*

La función *ll_angles* es quien calcula el gradiente de la imagen previamente filtrada para luego obtener el llamado *level-line orientation field*, en donde más tarde se hallarán los candidatos a segmentos. Lo que se hizo en esta función fué limitar el cálculo del gradiente a los píxeles donde la imagen escalada haya sido efectivamente computada. De esta manera se ahorra procesamiento innecesario, además de no detectarse las líneas negras en el contorno de la imagen (Figura 5.6), que de no ser así se detectarían.

5.6.3. Refinamiento y mejora de los candidatos

Se vió en la explicación del algoritmo el problema de que si hubiesen dos o más segmentos que formen entre ellos ángulos menores o iguales al valor umbral τ , estos serían detectados como uno único, heredado del algoritmo de Burns, Hanson y Riseman; y se explicó cómo, mediante un refinamiento de los segmentos, LSD soluciona este problema. Se vió además que luego de la validación o no de los segmentos previamente detectados, se realiza una mejora de los mismos para intentar que los no validados a causa de una mala estimación rectangular, sí puedan serlo.

Como en este proyecto en particular se trabaja con marcadores formados por cuadrados concéntricos, de bordes bien marcados y que forman ángulos rectos entre sí, el refinamiento y la mejora de

los candidatos no es algo que afecte la detección de los mismos; y por consiguiente se suprimieron ambos bloques. Como era de esperarse, dichas supresiones no significaron un cambio considerable en el algoritmo desde el punto de vista del desempeño ni del tiempo de ejecución cuando tan sólo se enfoca al marcador. Sin embargo, si las imágenes capturadas cuentan con muchos segmentos (imágenes naturales genéricas), se ve que la detección de los mismos es menos precisa que la del algoritmo original, pero que los tiempos de procesamiento son notablemente inferiores.

5.6.4. Algoritmo en precisión simple

Originalmente, LSD fue implementado en precisión doble o *double* (en general 64 bits por valor). Sin embargo, el *ipad 2* (dispositivo para el cual se optimizó el algoritmo), cuenta con un procesador *ARM Cortex-A9*, cuyo bus de datos es de 32 bits. Se decidió entonces probar cambiar al algoritmo a precisión simple o *float* (32 bits por valor) y los resultados fueron realmente buenos. No sólo el algoritmo bajó su tiempo de ejecución, sino que además no existen cambios notorios en el desempeño del mismo.

5.6.5. Resultados

5.6.5.1. Filtro Gaussiano



Figura 5.7: Imagen sintética del marcador trasladado y rotado.

Se analizaron los tiempos promedio para la ejecución del filtro Gaussiano original y del optimizado, ambos con precisión doble y simple. La imagen de prueba fue la de la Figura 5.7; sépase que por cómo es el algoritmo, el contenido de la imagen es independiente del tiempo de procesamiento en cualquiera de los casos, por lo que basta con una única imagen de prueba para sacar conclusiones respecto del desempeño del mismo. Los valores relevantes del experimento se muestran en las tablas 5.1 y 5.2:

- **Precisión doble (*double*)**
- **Precisión simple (*float*)**

5.6.5.2. Line Segment Detection

Se analizaron los tiempos conjuntos para la ejecución de LSD más el filtro Gaussiano, los originales y los optimizados, ambos con precisión doble y simple. Se probaron ambos bloques juntos ya que el algoritmo original está implementado con éstos integrados. Las imágenes de prueba fueron

	Filtro original	Filtro optimizado
Tamaño de imagen de entrada	480×360	480×360
Escala	0,5	0,5
Tamaño de imagen de salida	240×180	240×180
Segmentos detectados	36	36
Tiempo medio de procesamiento	36ms	29ms

Tabla 5.1: Comparación entre los tiempos de ejecución del filtro Gaussiano optimizado y el original. Ambos con precisión doble.

	Filtro original	Filtro optimizado
Tamaño de imagen de entrada	480×360	480×360
Escala	0,5	0,5
Tamaño de imagen de salida	240×180	240×180
Segmentos detectados	36	36
Tiempo medio de procesamiento	28ms	20ms

Tabla 5.2: Comparación entre los tiempos de ejecución del filtro Gaussiano optimizado y el original. Ambos con precisión simple.



Figura 5.8: Imagen *zebras.png*.

la del marcador sintético (Figura 5.7) y *zebras.png* mostrada en la Figura 5.8. Los valores relevantes de los experimentos se muestran en las tablas 5.3, 5.4, 5.5 y 5.6.

■ Precisión doble (*double*)

	Algoritmo original	Algoritmo optimizado
Imagen utilizada	marcador sintético	marcador sintético
Tamaño de imagen de entrada	480×360	480×360
Escala	0,5	0,5
Tamaño de imagen de salida	240×180	240×180
Segmentos detectados	36	36
Tiempo medio de procesamiento	55,4ms	48ms

Tabla 5.3: Comparación entre los tiempos de ejecución del filtro Gaussiano más LSD optimizados y los originales, para la imagen 5.7. En todos los casos con precisión doble.

	Algoritmo original	Algoritmo optimizado
Imagen utilizada	<i>zebras.png</i>	<i>zebras.png</i>
Tamaño de imagen de entrada	480×360	480×360
Escala	0,5	0,5
Tamaño de imagen de salida	240×180	240×180
Segmentos detectados	251	179
Tiempo medio de procesamiento	179,7ms	94,4ms

Tabla 5.4: Comparación entre los tiempos de ejecución del filtro Gaussiano más LSD optimizados y los originales, para la imagen 5.8. En todos los casos con precisión doble.

■ Precisión simple (*float*)

	Algoritmo original	Algoritmo optimizado
Imagen utilizada	marcador sintético	marcador sintético
Tamaño de imagen de entrada	480×360	480×360
Escala	0,5	0,5
Tamaño de imagen de salida	240×180	240×180
Segmentos detectados	36	36
Tiempo medio de procesamiento	47,8ms	38,8ms

Tabla 5.5: Comparación entre los tiempos de ejecución del filtro Gaussiano más LSD optimizados y los originales, para la imagen 5.7. En todos los casos con precisión simple.

	Algoritmo original	Algoritmo optimizado
Imagen utilizada	<i>zebras.png</i>	<i>zebras.png</i>
Tamaño de imagen de entrada	480×360	480×360
Escala	0,5	0,5
Tamaño de imagen de salida	240×180	240×180
Segmentos detectados	252	182
Tiempo medio de procesamiento	189,8ms	90,8ms

Tabla 5.6: Comparación entre los tiempos de ejecución del filtro Gaussiano más LSD optimizados y los originales, para la imagen 5.8. En todos los casos con precisión simple.

5.7. Conslusión

En el presente capítulo se vió en detalle LSD, un algoritmo de detección de segmentos en imágenes que puede ser considerado el estado del arte en su rubro. Luego se afirmó que, si bien sus autores sostienen que el algoritmo es temporalmente lineal, lo que haría viable su uso en tiempo real; la implementación disponible del mismo, realizada por los propios autores, no está optimizada para tal caso y por eso hubo que hacerle algunos pequeños cambios al código. Dichos cambios lograron mejoras importantes en cuanto al tiempo de procesamiento, manteniendo prácticamente invariado su desempeño.

Los resultados expuestos en las Tablas 5.1 y 5.2 pueden interpretarse como que la decisión de cambiar la precisión de la implementación del algoritmo de *double* a *float* fue una idea acertada. Por

su parte, la lectura de las Tablas 5.3, 5.4, 5.5 y 5.6 sugiere que las mejoras en los tiempos de LSD optimizado para el tiempo real, respecto del original, son del orden de:

30 % para imágenes con pocos segmentos;
50 % para imágenes naturales genéricas, con muchos segmentos.

Cabe aclarar sin embargo, que si bien los resultados cualitativos¹ sugieren resultados similares, los resultados cuantitativos fueron logrados con tan sólo las dos imágenes presentadas anteriormente.

Finalmente, en las Figuras 5.9 y 5.10 se muestran los resultados luego de haber procesado a las Figuras 5.7 y 5.8 con LSD, primero con la implementación original y luego con la optimizada. Se puede ver, en primer lugar, la gran diferencia que existe entre la cantidad de segmentos detectados en una y otra imagen. Además, se concluye que el desempeño de ambas implementaciones es muy similar para el caso de la imagen *zebras.png* e idéntico para el caso del marcador.



Figura 5.9: Resultado de procesar a la Figura 5.7 con LSD. Izq.: Implementación original. Der.: Implementación optimizada.

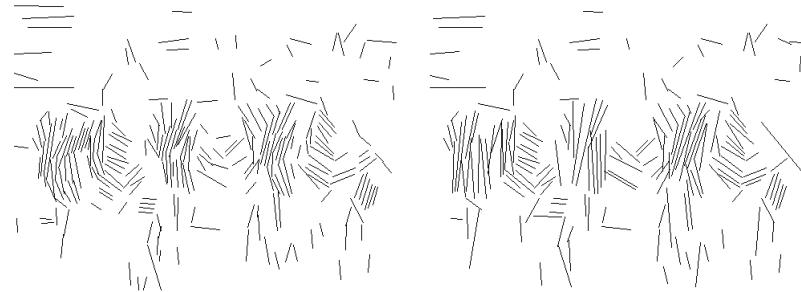


Figura 5.10: Resultado de procesar a la Figura 5.8 con LSD. Izq.: Implementación original. Der.: Implementación optimizada.

¹Se entiende por resultados cualitativos a la ejecución de LSD en tiempo real en el *iPad*, imposibles de ilustrar en el presente texto. Ver Sección 10.2.

CAPÍTULO 6

Modelo de cámara y estimación de pose monocular

6.1. Introducción

Se le llama “estimación de pose” al proceso mediante el cual se calcula en qué punto del mundo y con qué orientación se encuentra determinado objeto respecto de un eje de coordenadas previamente definido al que se lo llama “ejes del mundo”. Las aplicaciones de realidad aumentada requieren de un modelado preciso del entorno respecto de estos ejes, para poder ubicar correctamente los agregados virtuales dentro del modelo y luego dibujarlos de forma coherente en la imagen vista por el usuario. El objeto cuya estimación de pose resulta de mayor importancia es la cámara, ya que por ésta es por donde se mira la escena y es respecto de ésta que los objetos virtuales deben ubicarse de manera consistente. Una forma de estimar la pose de la cámara es mediante el uso de las imágenes capturadas por ella misma. Asimismo, el concepto “monocular” hace referencia al uso de una sola cámara, ya que es posible trabajar con más de una.

Para poder obtener información relevante a partir de las imágenes tomadas por una cámara, resulta necesario contar con un modelo preciso de su arquitectura ya que no todas las cámaras son iguales. El modelo más comúnmente utilizado es el denominado *pin-hole*. Para modelar completamente la arquitectura de la cámara se deben estimar ciertos “parámetros intrínsecos” a ésta, y eso se logra luego de realizados ciertos experimentos. A la estimación de estos parámetros se le denomina “calibración de la cámara”.

En este capítulo se verá en detalle el modelo de cámara *pin-hole*, tomando en cuenta la distorsión introducida por las lentes. Más adelante, se mencionarán distintos métodos para la calibración de una cámara y se verá en detalle en particular, el método de Zhang.

También se presentan los algoritmos más utilizados para el problema de estimación de pose, entre ellos el DLT(Direct Linear Transform), PnP(Perspective n Point) y RANSAC(RANdom SAmple Consensus).

Finalmente se presentan las diferentes maneras que hay para representar los ángulos de la pose y los problemas que se presentan cuando se trabaja con estas representaciones.

6.2. Modelo de cámara *pin-hole* [1]

6.2.1. Fundamentos y definiciones

Este modelo consiste en un centro óptico O , en donde convergen todos los rayos de la proyección y un plano imagen en el cual la imagen es proyectada. Se define *distancia focal* (f) como la distancia entre el centro óptico O y la intersección del eje óptico con el plano imagen (punto C). Ver Figura 6.1.

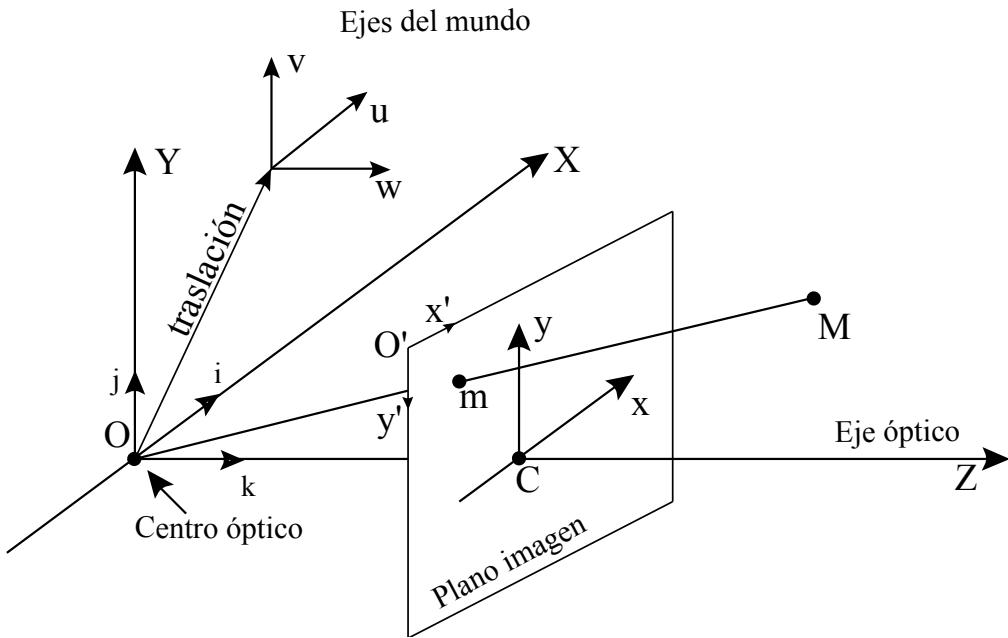


Figura 6.1: Modelo de cámara pin-hole.

Se llama proceso de proyección al proceso en el que se asocia al punto M del mundo, un punto \mathbf{m} en la imagen. Para modelar el mismo es necesario referirse a varias transformaciones y varios ejes de coordenadas.

- *Coordenadas del mundo:* son las coordenadas que describen la posición 3D del punto M respecto de los ejes del mundo (u, v, w). La elección de los ejes del mundo es arbitraria.
- *Coordenadas de la cámara:* son las coordenadas que describen la posición del punto M respecto de los ejes de la cámara (X, Y, Z). i, j y k son los versores de este eje de coordenadas.
- *Coordenadas de la imagen:* son las coordenadas que describen la posición del punto 2D, \mathbf{m} respecto del centro del plano imagen, C . Los ejes de este sistema de coordenadas son (x, y) .
- *Coordenadas normalizadas de la imagen:* son las coordenadas que describen la posición del punto 2D, \mathbf{m} , respecto del eje de coordenadas (x', y') situado en la esquina superior izquierda del plano imagen.

La transformación que lleva al punto M , expresado respecto de los ejes del mundo, al punto \mathbf{m} , expresado respecto del sistema de coordenadas normalizadas de la imagen, se puede ver como la composición de dos transformaciones menores. La primera, es la que realiza la proyección que transforma a un punto definido respecto del sistema de coordenadas de la cámara (X, Y, Z) en otro punto sobre el plano imagen expresado respecto del sistema de coordenadas normalizadas de la

imagen (x', y') . Véase que una vez calculada esta transformación, es una constante característica de cada cámara. Al conjunto de valores que definen esta transformación, se le llama “parámetros intrínsecos” de la cámara. La segunda, es la transformación que lleva de expresar un punto respecto de los ejes del mundo (u, v, w) , a ser expresado según los ejes de la cámara (X, Y, Z) . Esta última transformación varía conforme se mueve la cámara (respecto de los ejes del mundo) y el conjunto de valores que la definen es denominado “parámetros extrínsecos” de la cámara. Del cálculo de estos parámetros es que se obtiene la estimación de la pose de la cámara.

De lo anterior se concluye rápidamente que si se le llama H a la matriz proyección total, tal que:

$$\mathbf{m} = H \cdot \mathbf{M},$$

entonces:

$$H = I \cdot E$$

donde I corresponde a la matriz proyección asociada a los parámetros intrínsecos y E corresponde a la matriz asociada a los parámetros extrínsecos. Ambos juegos de parámetros acarrean información muy valiosa:

- **Parámetros extrínsecos:** pose de la cámara.

- Traslación: ubicación del centro óptico de la cámara respecto de los ejes del mundo.
- Rotación: rotación del sistema de coordenadas de la cámara (X, Y, Z) , respecto de los ejes del mundo.

- **Parámetros intrínsecos:** parámetros propios de la cámara. Dependen de su geometría interna y de su óptica.

- Punto principal ($\mathbf{C} = [x'_C, y'_C]$): es el punto intersección entre el eje óptico y el plano imagen. Las coordenadas de este punto vienen dadas en píxeles y son expresadas respecto del sistema normalizado de la imagen.
- Factores de conversión píxel-milímetros (d_x, d_y): indican el número de píxeles por milímetro que utiliza la cámara en las direcciones x e y respectivamente.
- Distancia focal (f): distancia entre el centro óptico (\mathbf{O}) y el punto principal (\mathbf{C}). Su unidad es el milímetro.
- Factor de proporción (s): indica la proporción entre las dimensiones horizontal y vertical de un píxel.

6.2.2. Matriz de proyección

En la sección anterior se vio que es posible hallar una “matriz de proyección” H que dependa tanto de los parámetros intrínsecos de la cámara como de sus parámetros extrínsecos:

$$\mathbf{m} = H \cdot \mathbf{M}$$

donde \mathbf{M} y \mathbf{m} son los puntos ya definidos y vienen expresados en “coordenadas homogéneas”. Por más información acerca de este tipo de coordenadas ver [8].

Para determinar la forma de la matriz de proyección se estudia cómo se relacionan las coordenadas de \mathbf{M} con las coordenadas de \mathbf{m} ; para hallar esta relación se debe analizar cada transformación, entre los sistemas de coordenadas mencionados con anterioridad, por separado.

- **Proyección 3D - 2D:** de las coordenadas homogéneas del punto **M** expresadas en el sistema de coordenadas de la cámara (X_0, Y_0, Z_0, T_0), a las coordenadas homogéneas del punto **m** expresadas en el sistema de coordenadas de la imagen (x_0, y_0, s_0):
Se desprende de la imagen 6.1 y algo de trigonometría la siguiente relación entre las coordenadas en cuestión y la distancia focal (f):

$$\frac{f}{Z_0} = \frac{x_0}{X_0} = \frac{y_0}{Y_0}$$

A partir de la relación anterior:

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \frac{f}{Z_0} \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix}$$

Expresado en forma matricial, en coordenadas homogéneas:

$$\begin{pmatrix} x_0 \\ y_0 \\ s_0 \end{pmatrix} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{pmatrix}$$

- **Transformación imagen - imagen:** de las coordenadas homogéneas del punto **m** expresadas respecto del sistema de coordenadas de la imagen (x_0, y_0, s_0), a las coordenadas homogéneas de él mismo pero expresadas respecto del sistema de coordenadas normalizadas de la imagen (x'_0, y'_0, s'_0):

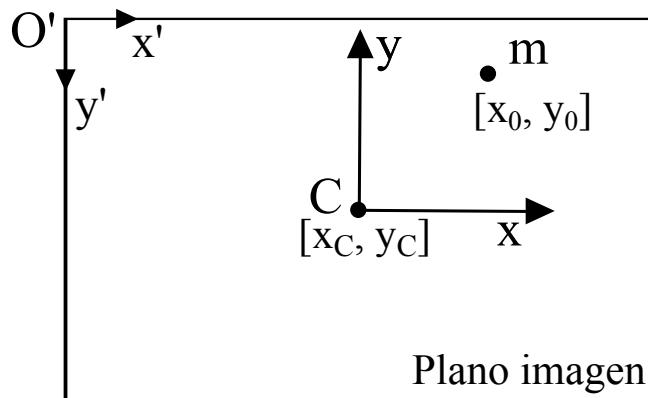


Figura 6.2: Relación entre el sistema de coordenadas de la imagen y el sistema de coordenadas normalizadas de la imagen.

Se les suma, a las coordenadas de **m** respecto del sistema de la imagen, la posición del punto **C** respecto del sistema normalizado de la imagen (x'_C, y'_C). Las coordenadas de **m** dejan de ser expresadas en milímetros para ser expresadas en píxeles. Aparecen los factores de conversión d_x y d_y :

$$\begin{aligned} x'_0 &= d_x \cdot x_0 + x'_C \\ y'_0 &= d_y \cdot y_0 + y'_C \end{aligned}$$

Se obtiene entonces la siguiente relación matricial, en coordenadas homogéneas:

$$\begin{pmatrix} x'_0 \\ y'_0 \\ s'_0 \end{pmatrix} = \begin{pmatrix} d_x & 0 & x'_C \\ 0 & d_y & y'_C \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \\ 1 \end{pmatrix}$$

- **Matriz de parámetros intrínsecos (I):** de las coordenadas homogéneas del punto **M** expresadas en el sistema de coordenadas de la cámara ($X_0, Y_0, Z_0, 1$), a las coordenadas homogéneas del punto **m** expresadas respecto del sistema de coordenadas normalizadas de la imagen (x'_0, y'_0, s'_0):

Se obtiene combinando las dos últimas transformaciones. Nótese que como ya se aclaró, depende únicamente de parámetros propios de la construcción de la cámara:

$$I = \begin{pmatrix} d_x \cdot f & 0 & x'_C & 0 \\ 0 & d_y \cdot f & y'_C & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Nota: De forma genérica se puede agregar a la matriz de parámetros intrínsecos del modelo *pin-hole* un parámetro s llamado en inglés *skew parameter*, o “parámetro de proporción” en Español. Este parámetro toma valores distintos de cero muy rara vez, pues modela los casos en los que los ejes x e y de los píxeles de la cámara no son perpendiculares entre sí. En casos realistas, $s \neq 0$ cuando por ejemplo se toma una fotografía de una fotografía. La matriz de parámetros intrínsecos, tomando en cuenta este parámetro, tendrá la forma:

$$I = \begin{pmatrix} d_x \cdot f & s & x'_C & 0 \\ 0 & d_y \cdot f & y'_C & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

- **Matriz de parámetros extrínsecos (E):** de las coordenadas homogéneas del punto **M** expresadas respecto del sistema de coordenadas del mundo (U_0, V_0, W_0, P_0), a las coordenadas homogéneas de él mismo pero expresadas respecto del sistema de coordenadas de la cámara (X_0, Y_0, Z_0, T_0):

Se obtiene de estimar la pose de la cámara respecto de los ejes del mundo y es la combinación de, primero una rotación R_{3x3} , y luego una traslación T_{3x1} . Se obtiene entonces la siguiente representación matricial:

$$\begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \\ T_0 \end{pmatrix} = \begin{pmatrix} R & T \\ 0 & 1 \end{pmatrix} \begin{pmatrix} U_0 \\ V_0 \\ W_0 \\ P_0 \end{pmatrix}$$

donde la matriz de parámetros extrínsecos desarrollada toma la forma:

$$E = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- **Matriz de proyección (H):** de las coordenadas homogéneas del punto **M** expresadas respecto del sistema de coordenadas del mundo (U_0, V_0, W_0, P_0), a las coordenadas homogéneas del punto **m** expresadas respecto del sistema de coordenadas normalizadas de la imagen (x'_0, y'_0, s'_0):

Es la proyección total y se obtiene combinando las dos transformaciones anteriores:

$$\begin{pmatrix} x'_0 \\ y'_0 \\ s'_0 \end{pmatrix} = \begin{pmatrix} d_x \cdot f & 0 & x'_C & 0 \\ 0 & d_y \cdot f & y'_C & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} U_0 \\ V_0 \\ W_0 \\ P_0 \end{pmatrix}$$

6.3. Distorsión introducida por las lentes

Hasta el momento se asumió que el modelo lineal presentado para la proyección de cualquier punto del mundo en el plano imagen de la cámara es lo suficientemente preciso en todos los casos. Sin embargo, en casos reales, y cuando las lentes de las cámaras no son del todo buenas, la distorsión introducida por estas se hace notar. Dado el punto \mathbf{M} de coordenadas (X_0, Y_0, Z_0) respecto de los ejes de la cámara, se le llama distorsión a la diferencia entre su proyección ideal en el plano imagen (x_0, y_0) y su proyección real $(\tilde{x}_0, \tilde{y}_0)$. La más común de todas, es la denominada “distorsión radial”, ya que su magnitud depende del radio medido desde el punto principal del plano imagen, hasta las coordenadas del punto en cuestión.

La forma de solucionar el presente problema es realizar una corrección de la distorsión, modelando a la misma de la siguiente manera:

$$\begin{pmatrix} \tilde{x}_0 \\ \tilde{y}_0 \end{pmatrix} = L(r) \cdot \begin{pmatrix} x_0 \\ y_0 \end{pmatrix},$$

donde r es la distancia radial $\sqrt{x_0^2 + y_0^2}$ y $L(r)$ es un factor de distorsión que depende únicamente del radio r . Si se desarrolla la ecuación anterior, y se expresa en píxeles, respecto del sistema de coordenadas normalizadas de la imagen; se obtiene lo siguiente:

$$\begin{aligned} \tilde{x}'_0 &= x'_C + L(r)(x'_0 - x'_C) \\ \tilde{y}'_0 &= y'_C + L(r)(y'_0 - y'_C) \end{aligned}$$

donde $(\tilde{x}'_0, \tilde{y}'_0)$ son las coordenadas reales de la proyección medidas en píxeles, (x'_0, y'_0) son las coordenadas ideales de la proyección medidas también en píxeles y (x'_C, y'_C) son las coordenadas del punto principal. Véase que en este caso $r = \sqrt{(x'_0 - x'_C)^2 + (y'_0 - y'_C)^2}$.

La función $L(r)$ es definida sólo para valores positivos de r y $L(0) = 1$. Una aproximación a la función arbitraria $L(r)$ puede ser una expansión de Taylor: $L(r) = 1 + k_1 r + k_2 r^2 + k_3 r^3 + \dots$. Finalmente, a la hora de calcular los parámetros intrínsecos de una cámara, también deben ser estimados sus coeficientes de distorsión radial $\{k_1, k_2, k_3, k_4, \dots\}$.

6.4. Métodos para la calibración de cámara

Como se vio algunos párrafos atrás, el proceso mediante el cual se calculan los parámetros intrínsecos reales de una cámara es denominado “calibración de cámara”. Existen varios métodos para calibrar una cámara; sin embargo, los tres algoritmos, basados en modelos planos, más ampliamente utilizados alrededor del mundo [?] son el método de Zhang [?], el método de R.Y. Tsai [?] y un método llamado “Direct Linear Transform” (DLT) [?]. Para calibrar las cámaras utilizadas en este proyecto, se trabajó con una implementación en *Matlab* basada en el método de Zhang ([?]), que afortunadamente dio resultados muy buenos. Por eso, se explicará a continuación, de forma breve, cómo funciona este método. Por dudas respecto de cualquier resultado matemático expuesto sin los cálculos intermedios, siempre se recomienda leer el artículo original.

El método de Zhang es muy sencillo y flexible. Sólo requiere de la cámara a calibrar, una computadora y una imagen patrón (plana), de tipo damero; a la que se le tomarán al menos dos fotografías desde orientaciones distintas. En la figura 6.3 se ve una de las imágenes utilizadas para



Figura 6.3: Imagen de un damero, utilizada para calibrar la cámara del *iPad* durante el proyecto.

calibrar la cámara del *iPad* durante el proyecto. Ni las posiciones de la cámara en cada caso, ni el movimiento entre estas posiciones tienen por qué ser conocidos. Este método devuelve los parámetros intrínsecos de la cámara correspondientes al modelo *pin-hole* visto anteriormente, sus parámetros extrínsecos para cada fotografía utilizada para la calibración y la distorsión radial de sus lentes.

Recuérdese que la relación entre un punto 3D \mathbf{M} expresado respecto de los ejes de coordenadas del mundo y su proyección en el plano imagen \mathbf{m} , expresada respecto de los ejes normalizados de la imagen, viene dada por:

$$\mathbf{m} = \mathbf{I} \cdot \mathbf{E} \cdot \mathbf{M}$$

donde \mathbf{E} representa a la matriz de parámetros extrínsecos e \mathbf{I} representa a la matriz de parámetros intrínsecos de la cámara. Además:

$$\mathbf{I} = \begin{pmatrix} \alpha & s & x'_C \\ 0 & \beta & y'_C \\ 0 & 0 & 1 \end{pmatrix}$$

con $\alpha = d_x \cdot f$ y $\beta = d_y \cdot f$.

Se asume en este método que el sistema de coordenadas del mundo “reposa” sobre la imagen patrón; o lo que es lo mismo, que esta se encuentra en $Z = 0$. Se obtiene entonces la siguiente simplificación:

$$\begin{pmatrix} x'_0 \\ y'_0 \\ 1 \end{pmatrix} = \mathbf{I} \cdot \begin{pmatrix} r_1 & r_2 & r_3 & t \end{pmatrix} \cdot \begin{pmatrix} U_0 \\ V_0 \\ W_0 \\ 1 \end{pmatrix} = \mathbf{I} \cdot \begin{pmatrix} r_1 & r_2 & t \end{pmatrix} \cdot \begin{pmatrix} U_0 \\ V_0 \\ 1 \end{pmatrix}$$

donde $(U_0, V_0, W_0, 1)^T$ denota las coordenadas homogéneas del punto \mathbf{M} respecto de los ejes del mundo y $(x'_0, y'_0, 1)^T$ representa las coordenadas homogéneas de su proyección en el plano imagen, \mathbf{m} , respecto de los ejes normalizados de la imagen. Se le llamó r_i a la i-ésima columna de la matriz rotación de los parámetros extrínsecos de la cámara.

Dada una fotografía de la imagen patrón plana (figura 6.3), es posible estimar una homografía que relacione a los puntos de la imagen con sus correspondientes en la fotografía. Si se toma en cuenta que dicha homografía vale $H = (h_1, h_2, h_3) = \mathbf{I} \cdot (r_1, r_2, t)$, con h_i la i-ésima columna de la matriz, y que las columnas r_1 y r_2 son ortonormales entre sí, realizando algo de matemática se llega

a que:

$$\begin{aligned} h_1^T \cdot (I^{-1})^T \cdot I^{-1} \cdot h_2 &= 0 \\ h_1^T \cdot (I^{-1})^T \cdot I^{-1} \cdot h_1 &= h_2^T \cdot (I^{-1})^T \cdot I^{-1} \cdot h_2 \end{aligned}$$

Las anteriores son las únicas dos relaciones básicas entre parámetros intrínsecos que se pueden obtener a partir de una única homografía. Esto es porque una homografía tiene 8 grados de libertad y existen 6 parámetros extrínsecos (3 para la traslación y 3 para la rotación).

Si se define la matriz B como sigue:

$$B = (I^{-1})^T \cdot I^{-1} = \begin{pmatrix} B_{11} & B_{21} & B_{31} \\ B_{12} & B_{22} & B_{32} \\ B_{13} & B_{23} & B_{33} \end{pmatrix} = \begin{pmatrix} \frac{1}{\alpha^2} & -\frac{s}{\alpha^2 \cdot \beta} & \frac{s \cdot v'_P - u'_P \cdot \beta}{\alpha^2 \cdot \beta} \\ -\frac{s}{\alpha^2 \cdot \beta} & \frac{s^2}{\alpha^2 \cdot \beta^2} + \frac{1}{\beta^2} & -\frac{s(s \cdot v'_P - u'_P \cdot \beta)}{\alpha^2 \cdot \beta^2} - \frac{v'_P}{\beta^2} \\ \frac{s \cdot v'_P - u'_P \cdot \beta}{\alpha^2 \cdot \beta} & -\frac{s(s \cdot v'_P - u'_P \cdot \beta)}{\alpha^2 \cdot \beta^2} - \frac{v'_P}{\beta^2} & \frac{(s \cdot v'_P - u'_P \cdot \beta)^2}{\alpha^2 \cdot \beta^2} + \frac{v'^2_P}{\beta^2} + 1 \end{pmatrix}$$

se ve fácilmente que esta es simétrica, por lo que quedará absolutamente definida por un vector de 6 dimensiones:

$$b = (B_{11}, B_{12}, B_{22}, B_{13}, B_{23}, B_{33})^T$$

Si además se define el vector variable v_{ij} de la siguiente manera:

$$v_{ij} = (h_{i1} \cdot h_{j1}, h_{i1} \cdot h_{j2} + h_{i2} \cdot h_{j1}, h_{i2} \cdot h_{j2}, h_{i3} \cdot h_{j1} + h_{i1} \cdot h_{j3}, h_{i3} \cdot h_{j2} + h_{i2} \cdot h_{j3}, h_{i3} \cdot h_{j3})^T,$$

se tiene que:

$$h_i^T \cdot B \cdot h_j = V_{ij}^T \cdot b$$

Las dos relaciones básicas entre parámetros intrínsecos obtenidas de una única homografía, vistas anteriormente, pueden ser reescritas como:

$$\begin{pmatrix} v_{12}^T \\ (v_{11} - v_{22})^T \end{pmatrix} \cdot b = V \cdot b = 0$$

Utilizando n fotografías distintas de la imagen patrón, y por lo tanto n homografías distintas se obtiene una matriz V de tamaño $2 \cdot n \times 6$. Es sabido que si $n \geq 3$, el sistema matricial anterior tendrá una solución b única, que varía según cierto factor de escala. Sin embargo, si $n = 2$, es posible imponer la condición $s = 0$ y así también calcular al vector b de forma única, sin mayores problemas.

Una vez estimado b es posible reconstruir la matriz de parámetros intrínsecos I , para luego utilizando I y las homografías H obtener los parámetros extrínsecos de la cámara para cada fotografía utilizada para la calibración.

El artículo de Zhang afirma que la solución obtenida hasta el momento no es del todo buena, pues se obtuvo minimizando una distancia algebraica y eso no tiene mucho sentido. Lo que se hace entonces es, utilizando las n fotografías tomadas para la calibración y los k puntos seleccionados en cada una de ellas, minimizar la siguiente ecuación:

$$\sum_{i=1}^n \sum_{j=1}^k \|m_{ij} - \hat{m}(I, E_i, M_j)\|^2$$

donde $\hat{m}(I, E_i, M_j)$ es la proyección del punto M_j en la imagen i utilizando la homografía $H_i = I \cdot E_i$. El resultado de dicha minimización no lineal será el resultado final. Este método requiere de valores iniciales para I y para los $E_i|_{i=1..n}$; que serán los obtenidos en los cálculos anteriores.

Finalmente se realiza una estimación de la distorsión radial utilizando un modelo muy similar al visto en la sección 6.3.

6.5. Calibración de cámaras utilizadas

A lo largo del proyecto se utilizaron diferentes dispositivos, se utilizó un *iPad 2*, un *iPhone 4* y un *iPod Touch 4^{ta} generación*. Todas las cámaras fueron calibradas con el método de Zhang. La distorsión radial no se tomó en cuenta para ningún caso, aún así los resultados obtenidos buenos. Se recuerda que la matriz de parámetros intrínsecos se expresa como

$$I = \begin{pmatrix} d_x \cdot f & s & x'_C \\ 0 & d_y \cdot f & y'_C \\ 0 & 0 & 1 \end{pmatrix}$$

A continuación se muestran las calibraciones, todos los valores están expresados en píxeles.

6.5.1. Calibración *iPod Touch 4^{ta} generación*

Se realizaron dos calibraciones, la primera calibración que se realizó fue para imágenes de 640×480 , la matriz obtenida fue:

$$I = \begin{pmatrix} 746,36170 & 0 & 292,80331 \\ 0 & 745,43429 & 217,56288 \\ 0 & 0 & 1 \end{pmatrix}$$

La segunda calibración fue para imágenes de 960×720 . Se obtuvo la siguiente matriz:

$$I = \begin{pmatrix} 1154,46114 & 0 & 473,51550 \\ 0 & 1152,02675 & 418,95929 \\ 0 & 0 & 1 \end{pmatrix}$$

6.5.2. Calibración *iPhone 4*

Se realizaron dos calibraciones, la primera calibración fue para imágenes de 640×480 , se obtuvieron los siguientes parámetros intrínsecos:

$$I = \begin{pmatrix} 621,54488 & 0 & 345,63801 \\ 0 & 617,33033 & 235,04564 \\ 0 & 0 & 1 \end{pmatrix}$$

La segunda calibración fue para imágenes de 352×288 , la matriz obtenida fue:

$$I = \begin{pmatrix} 355,72881 & 0 & 222,56769 \\ 0 & 348,80501 & 139,91097 \\ 0 & 0 & 1 \end{pmatrix}$$

6.5.3. Calibración *iPad 2*

Los parámetros intrínsecos obtenidos para imágenes de 480×360 son:

$$I = \begin{pmatrix} 589,141 & 0 & 205,115 \\ 0 & 580,754 & 165,912 \\ 0 & 0 & 1 \end{pmatrix}$$

6.6. Problema de estimación de pose

Como se mencionó en 6.2.1 la matriz de parámetros extrínsecos E representa a la pose de la cámara. El problema de estimación de pose consiste en determinar esta matriz dadas n correspondencias entre puntos M_i en el mundo 3D y puntos m_i en la imagen.

Existen varios algoritmos de estimación de pose, a continuación se presentan algunos.

6.6.1. DLT(Direct Linear Transform)

Este método sirve para calcular la matriz \mathbf{H} en la cual están implícitos los parámetros intrínsecos y extrínsecos. Si se conocen los parámetros intrínsecos se pueden despejar la matriz \mathbf{E} con la información de la pose.

Como se vio anteriormente

$$\begin{pmatrix} x_i \\ y_i \\ s_i \end{pmatrix} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \mathbf{H}_{13} & \mathbf{H}_{14} \\ \mathbf{H}_{21} & \mathbf{H}_{22} & \mathbf{H}_{23} & \mathbf{H}_{24} \\ \mathbf{H}_{31} & \mathbf{H}_{32} & \mathbf{H}_{33} & \mathbf{H}_{34} \\ \mathbf{H}_{41} & \mathbf{H}_{42} & \mathbf{H}_{43} & \mathbf{H}_{44} \end{pmatrix} \cdot \begin{pmatrix} U_i \\ V_i \\ W_i \\ P_i \end{pmatrix}$$

Por comodidad a partir de ahora cuando se refiera a puntos 2D (x_i, y_i) serán expresados siempre desde el eje de coordenadas normalizadas de la imagen. Se debe notar que la matriz \mathbf{H} puede ser multiplicada por un factor distinto de cero sin alterar el resultado de la proyección, esto se debe a que se trabaja con coordenadas homogéneas. Por lo tanto lo que define la proyección no son los elementos de \mathbf{H} sino la relación entre todos los elementos (excepto el de factor de escala) y el elemento que da el factor de escala. Así entonces \mathbf{H} tiene 12 elementos pero solamente 11 grados de libertad.

Cada correspondencia $M_i \leftrightarrow m_i$ aporta dos ecuaciones linealmente independientes con los elementos de la matriz \mathbf{H} , \mathbf{H}_{ij} como variables

$$\frac{\mathbf{H}_{11}X_i + \mathbf{H}_{12}Y_i + \mathbf{H}_{13}Z_i + \mathbf{H}_{14}}{\mathbf{H}_{31}X_i + \mathbf{H}_{32}Y_i + \mathbf{H}_{33}Z_i + \mathbf{H}_{34}} = x_i,$$

$$\frac{\mathbf{H}_{21}X_i + \mathbf{H}_{22}Y_i + \mathbf{H}_{23}Z_i + \mathbf{H}_{24}}{\mathbf{H}_{31}X_i + \mathbf{H}_{32}Y_i + \mathbf{H}_{33}Z_i + \mathbf{H}_{34}} = y_i$$

La ecuación para s_i se puede obtener como combinación lineal de las de x_i y y_i , por eso no se tiene en cuenta. Estas ecuaciones se pueden reescribir como $\mathbf{A}h = 0$ donde

$$h = (\mathbf{H}_{11} \quad \mathbf{H}_{12} \quad \mathbf{H}_{13} \quad \mathbf{H}_{14} \quad \mathbf{H}_{21} \quad \mathbf{H}_{22} \quad \mathbf{H}_{23} \quad \mathbf{H}_{24} \quad \mathbf{H}_{31} \quad \mathbf{H}_{32} \quad \mathbf{H}_{33} \quad \mathbf{H}_{34})^T$$

y

$$\mathbf{A} = \begin{pmatrix} X_0 & Y_0 & Z_0 & 1 & 0 & 0 & 0 & 0 & -x_0X_0 & -x_0Y_0 & -x_0Z_0 & -x_0 \\ 0 & 0 & 0 & 0 & X_0 & Y_0 & Z_0 & 1 & -y_0X_0 & -y_0Y_0 & -y_0Z_0 & -y_0 \\ X_1 & Y_1 & Z_1 & 1 & 0 & 0 & 0 & 0 & -x_1X_1 & -x_1Y_1 & -x_1Z_1 & -x_1 \\ 0 & 0 & 0 & 0 & X_1 & Y_1 & Z_1 & 1 & -y_1X_1 & -y_1Y_1 & -y_1Z_1 & -y_1 \\ \vdots & \vdots \\ X_{n-1} & Y_{n-1} & Z_{n-1} & 1 & 0 & 0 & 0 & 0 & -x_{n-1}X_{n-1} & -x_{n-1}Y_{n-1} & -x_{n-1}Z_{n-1} & -x_{n-1} \\ 0 & 0 & 0 & 0 & X_{n-1} & Y_{n-1} & Z_{n-1} & 1 & -y_{n-1}X_{n-1} & -y_{n-1}Y_{n-1} & -y_{n-1}Z_{n-1} & -y_{n-1} \end{pmatrix}$$

La transformación obtenida la matriz \mathbf{H} tiene 11 grados de libertad como mencionó anteriormente, por lo tanto el rango de la matriz \mathbf{A} es 11. Se realiza la descomposición SVD de la matriz \mathbf{A} , el

vector propio del valor singular de \mathbf{A} con valor menor es la base del núcleo de \mathbf{A} . Entonces se tiene que h es este vector a menos de una constante. Una vez que se tiene h se puede armar la matriz \mathbf{H} . Luego se tiene que

$$\mathbf{E} = \mathbf{I}^{-1} \mathbf{H}$$

6.6.2. *PnP (Perspective-n-Point)*

Si se cuenta con los parámetros intrínsecos, se puede utilizar un enfoque que se centre en calcular solamente la pose de la cámara. Dependiendo de la cantidad de correspondencias que se tienen entre la imagen y el modelo es posible obtener un número finito de soluciones de la pose. Si se tienen 1 o 2 correspondencias el problema tiene infinitas soluciones. Si se tienen 3 correspondencias(P3P) se obtienen hasta 4 posibles soluciones. Para 4 o más correspondencias se obtiene una única solución, siempre que los puntos no estén alineados. La idea detrás de este algoritmo es la siguiente:

- A partir de los puntos de la imagen m_i y conociendo la distancia focal f es posible calcular los versores j_i .

$$j_i = \frac{1}{\sqrt{x_i^2 + y_i^2 + f^2}} \begin{pmatrix} x_i \\ y_i \\ f \end{pmatrix}$$

- Con estos versores es posible determinar los ángulos que forman las líneas de vista de los puntos \mathbf{M}_i entre sí.
- Se busca estimar las distancias $l_i = \|\mathbf{OM}_i\|$ entre el centro de la cámara y los puntos 3D \mathbf{M}_i a partir de las relaciones dadas por los triángulos $\mathbf{OM}_i\mathbf{M}_j$.
- Una vez que se calculan las distancias l_i , los puntos \mathbf{M}_i se expresan en el sistema de coordenadas de la cámara como \mathbf{M}_i^C .
- Finalmente \mathbf{R} y \mathbf{T} quedan determinadas como la transformación que lleva puntos en el sistema de coordenadas del mundo a el sistema de coordenadas de la cámara.

En la bibliografía [?] y [?] se encuentran varios métodos para resolver numéricamente el problema.

6.6.3. **RANSAC(RANdom SAmple Consensus)**

Este es un algoritmo iterativo utilizado para estimar los parámetros de un modelo matemático de un conjunto de datos que contiene *outliers* (datos fuera del modelo). En particular se puede utilizar para el problema de estimación de pose cuando no se tienen las correspondencias entre puntos detectados y puntos del modelo.

A continuación se presenta el algoritmo:

- (1) Dado un modelo que requiere un mínimo de n puntos para determinar sus parámetros, y un conjunto de datos \mathbf{P} tal que el número de puntos en \mathbf{P} es mayor que n , se sortea un subconjunto S_1 de n puntos de \mathbf{P} para instanciar el modelo. Con el modelo instanciado \mathbf{M}_1 se determina el subconjunto de decisión S_1^* de puntos de \mathbf{P} que están a menos de una distancia t de \mathbf{M}_1 .
- (2) Si la cantidad de puntos en S_1^* es mayor que un umbral \mathbf{T} entonces se elige el subconjunto de decisión S_1^* para computar el nuevo modelo \mathbf{M}_1^* .

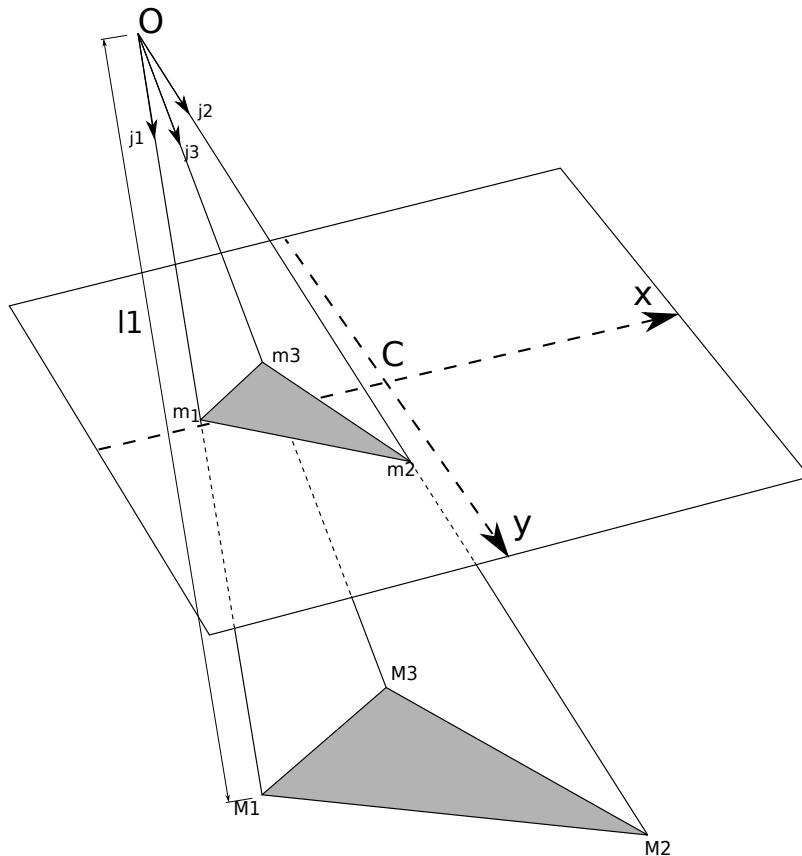


Figura 6.4: Geometría del problema P3P. Se busca calcular la distancia entre el centro óptico O y los puntos del modelo 3D

- (3) Si la cantidad de puntos en S_1^* es menor que \mathbf{T} , se sortea un nuevo subconjunto S_2 y se repite el proceso. Si luego de una cantidad de N número de pruebas no se obtiene un subconjunto de decisión que cumple con el umbral \mathbf{T} , se resuelve el modelo con el subconjunto de decisión mas grande obtenido, o se termina sin devolver modelo.

Los parámetros t , \mathbf{T} y N , se eligen en base al modelo a estimar y a la probabilidad de encontrar un *outlier* en el conjunto de datos .

6.6.4. POSIT

Este es un algoritmo iterativo que se basa en utilizar la proyección ortogonal escalada (SOP) para resolver el problema de estimación de pose. Se necesita tener más de cuatro correspondencias entre puntos del modelo M_i y puntos en la imagen m_i . De todos los algoritmos presentados este fue el que se decidió utilizar. El desarrollo de la teoría de este algoritmo se encuentra en 7. Este algoritmo tiene diferentes variantes. Por un lado esta la versión original del algoritmo y una versión que resuelve el caso en que todos los puntos del modelo están en un mismo plano, (POSIT Coplanar). Luego se tiene una variante llamada SoftPOSIT que resuelve la estimación de pose sin la necesidad de conocer las correspondencias entre puntos del modelo 3D y puntos de la imagen en el caso en que los puntos del modelo no sean coplanares. Finalmente se tiene una variante de SoftPOSIT que trabaja con líneas.

La variante de SoftPOSIT de líneas fue el principal argumento para tomar la decisión ya que el detector de características que se usa es el LSD refch: lsd. Esta variante fue implementada sin éxito, pero en busca de esta implementación se desarrolló una versión de POSIT para puntos coplanares que no está presentada en la bibliografía y dio buenos resultados. Otro argumento a favor de esta

opción es que se contaba con implemetaciones de algunas variantes. De [?] se obtuvieron las implementaciones de POSIT y POSIT coplanar en C y la implementación en MatLab de SoftPOSIT. Para la variante de SoftPOSIT de líneas sólo se contó con el artículo[?].

6.7. Representación de la pose de la cámara

Como se vio anteriormente, la pose de la cámara queda determinada por una matriz de rotación \mathbf{R} y un vector de traslación \mathbf{T} . La matriz \mathbf{R} indica la orientación de la cámara respecto al mundo. Hay varias maneras de representar esta orientación, entre ellas se encuentran la representación matricial, la representación en ángulos de Euler y los *quaternions*. Dependiendo de la aplicación puede resultar más útil utilizar las diferentes representaciones.

6.7.1. Representación matricial

Esta representación es la que se introdujo en 6.2.2. En esta matriz las filas corresponden a los versores del sistema de coordenadas de la cámara expresados en las coordenadas del mundo. Se puede expresar como

$$\mathbf{R} = \begin{pmatrix} i_u & i_v & i_w \\ j_u & j_v & j_w \\ k_u & k_v & k_w \end{pmatrix}$$

La ventaja que tiene esta representación es que el pasaje de puntos en coordenadas del mundo a coordenadas de la cámara es directo, simplemente se multiplica por la matriz \mathbf{R} al punto en coordenadas del mundo y se le suma el vector de traslación.

6.7.2. Ángulos de Euler

La matriz de rotación \mathbf{R} se puede escribir como un producto de matrices que representan las rotaciones alrededor de los ejes x , y y z . No hay ninguna convención establecida en cuanto al orden en que se realizan las rotaciones. Por ejemplo si se toman ψ , θ y ϕ como los ángulos de rotación en torno a x , y y z respectivamente se tiene

$$\mathbf{R} = R_z(\phi)R_y(\theta)R_x(\psi) = \begin{pmatrix} \cos \phi & -\sin \phi & 0 \\ \sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \psi & -\sin \psi \\ 0 & \sin \psi & \cos \psi \end{pmatrix}$$

Desarrollando el producto se tiene que

$$\mathbf{R} = \begin{pmatrix} \cos \theta \cos \phi & \sin \psi \sin \theta \cos \phi - \cos \psi \cos \phi & \cos \psi \sin \theta \cos \phi + \sin \psi \sin \phi \\ \cos \theta \sin \phi & \sin \psi \sin \theta \sin \phi + \cos \psi \cos \phi & \cos \psi \sin \theta \sin \phi - \sin \psi \cos \phi \\ -\sin \theta & \sin \psi \cos \theta & \cos \psi \cos \theta \end{pmatrix}$$

6.7.2.1. Orden de rotaciones

Cuando se trabaja con matrices de rotaciones y ángulos de Euler es necesario saber en qué orden se aplican las rotaciones, pues no es una transformación comutativa. Un objeto rotado primero según el eje x y luego según el eje y termina en una posición diferente que si se lo rota primero según y y luego según x . Esto se puede ver en la Figura 6.7.2.1.

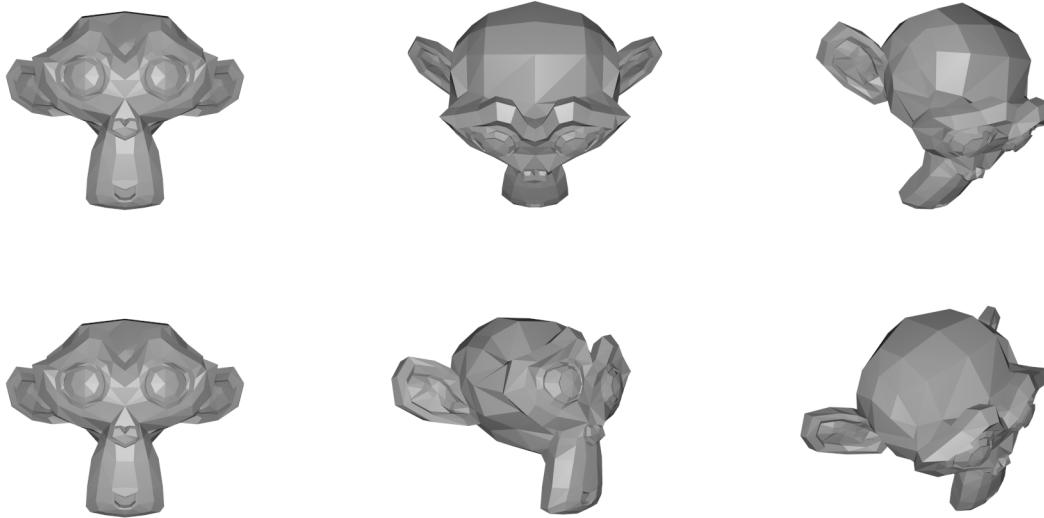


Figura 6.5: Se parte del mismo objeto, en la fila superior se aplica una rotación de 45° según x y luego una rotación de 45° según y . En la fila de abajo se aplican las mismas rotaciones pero en orden inverso. Se puede ver que se obtienen diferentes posiciones.

6.7.2.2. Cálculo de los ángulos de Euler

Si se tiene la matriz \mathbf{R} es posible realizar la descomposición y obtener los ángulos de Euler. De \mathbf{R}_{31} se obtiene el valor de θ

$$\theta = -\arcsin(\mathbf{R}_{31}).$$

Como $\sin(\theta) = \sin(\pi - \theta)$, puede haber dos posibles valores de θ (si $\mathbf{R}_{31} \neq \pm 1$)

$$\begin{aligned}\theta_1 &= -\arcsin(\mathbf{R}_{31}) \\ \theta_2 &= \pi - \theta_1 = \pi + \arcsin(\mathbf{R}_{31})\end{aligned}\tag{6.1}$$

A partir de estos valores de θ es posible encontrar dos juegos de ángulos que dan la misma matriz \mathbf{R} . Para calcular ψ se observa que

$$\frac{\mathbf{R}_{32}}{\mathbf{R}_{33}} = \tan \psi$$

de donde se deduce que

$$\psi = \arctan \left(\frac{\mathbf{R}_{32}}{\mathbf{R}_{33}} \right)$$

Es importante obtener el cuadrante al que pertenece el ángulo, por esto es que se usa la función $\arctan 2$ que esta disponible en C , recordar que la imagen de \arctan es $[-\pi/2, \pi/2]$ por lo que hay 2 cuadrantes que no se consideran. Como en los términos \mathbf{R}_{32} y \mathbf{R}_{33} aparece el término $\cos \theta$ multiplicando, hay que tener en cuenta su signo para obtener el valor de ψ . Si $\cos(\theta) > 0$, se tiene que $\psi = \arctan(\mathbf{R}_{32}/\mathbf{R}_{33})$. Si $\cos(\theta) < 0$, $\psi = \arctan(-\mathbf{R}_{32}/-\mathbf{R}_{33})$. Para tener en cuenta esto se toma

$$\psi = \arctan \left(\frac{\mathbf{R}_{32}/\cos \theta}{\mathbf{R}_{33}/\cos \theta} \right)$$

Por lo tanto los dos posibles valores para ψ son

$$\begin{aligned}\psi_1 &= \arctan\left(\frac{\mathbf{R}_{32}/\cos\theta_1}{\mathbf{R}_{33}/\cos\theta_1}\right) \\ \psi_2 &= \arctan\left(\frac{\mathbf{R}_{32}/\cos\theta_2}{\mathbf{R}_{33}/\cos\theta_2}\right)\end{aligned}\tag{6.2}$$

De manera similar se puede obtener ϕ . Se observa que

$$\frac{\mathbf{R}_{21}}{\mathbf{R}_{11}} = \tan\phi$$

por lo tanto se llega a

$$\begin{aligned}\phi_1 &= \arctan\left(\frac{\mathbf{R}_{21}/\cos\theta_1}{\mathbf{R}_{11}/\cos\theta_1}\right) \\ \phi_2 &= \arctan\left(\frac{\mathbf{R}_{21}/\cos\theta_2}{\mathbf{R}_{11}/\cos\theta_2}\right)\end{aligned}\tag{6.3}$$

Las ecuaciones 10.2 y 10.3 son válidas para el caso en que $\cos\theta \neq 0$

En el caso en que $\cos\theta = 0$ se tiene que $\theta = \pm\pi/2$, ademas los términos \mathbf{R}_{11} , \mathbf{R}_{21} , \mathbf{R}_{32} y \mathbf{R}_{33} son nulos. Por lo tanto se utilizan otros elementos de la matriz de rotación para hallar los ángulos restantes. En el caso en que $\theta = \pi/2$ se tiene que

$$\begin{aligned}\mathbf{R}_{12} &= \sin\psi\cos\phi - \cos\psi\cos\phi = \sin(\psi - \phi) \\ \mathbf{R}_{13} &= \cos\psi\cos\phi + \sin\psi\sin\phi = \cos(\psi - \phi) \\ \mathbf{R}_{22} &= \sin\psi\sin\phi + \cos\psi\cos\phi = \cos(\psi - \phi) = \mathbf{R}_{13} \\ \mathbf{R}_{23} &= \cos\psi\sin\phi - \sin\psi\cos\phi = -\sin(\psi - \phi) = -\mathbf{R}_{12}\end{aligned}$$

Cualquier ψ y ϕ que verifiquen estas ecuaciones serán soluciones válidas. Usando las ecuaciones para \mathbf{R}_{12} y \mathbf{R}_{13} se tiene que

$$\begin{aligned}\psi - \phi &= \arctan(\mathbf{R}_{12}/\mathbf{R}_{13}) \\ \psi &= \phi + \arctan(\mathbf{R}_{12}/\mathbf{R}_{13})\end{aligned}$$

Para el caso en que $\theta = -\pi/2$ se procede de igual manera y se llega a que

$$\begin{aligned}\psi + \phi &= \arctan(-\mathbf{R}_{12}/-\mathbf{R}_{13}) \\ \psi &= -\phi + \arctan(-\mathbf{R}_{12}/-\mathbf{R}_{13})\end{aligned}$$

6.7.2.3. Gimbal lock

La gran desventaja que presenta la representación de la rotación mediante ángulos de Euler, es el problema denominado *gimbal lock*. Este problema se da cuando 2 de los ejes de rotación quedan alineados. Si hay dos ejes alineados, se pierde un grado de libertad ya que los dos ejes rotan de la misma manera. Para la composición de rotaciones que se utilizan en la aplicación el *gimbal lock* se da cuando se gira $\pi/2$ según y . Como se vio anteriormente, en el caso en que $\theta = \pi/2$ la matriz de rotación queda

$$\mathbf{R} = \begin{pmatrix} 0 & \sin(\psi - \phi) & \cos(\psi - \phi) \\ 0 & \cos(\psi - \phi) & -\sin(\psi - \phi) \\ -1 & 0 & 0 \end{pmatrix}$$

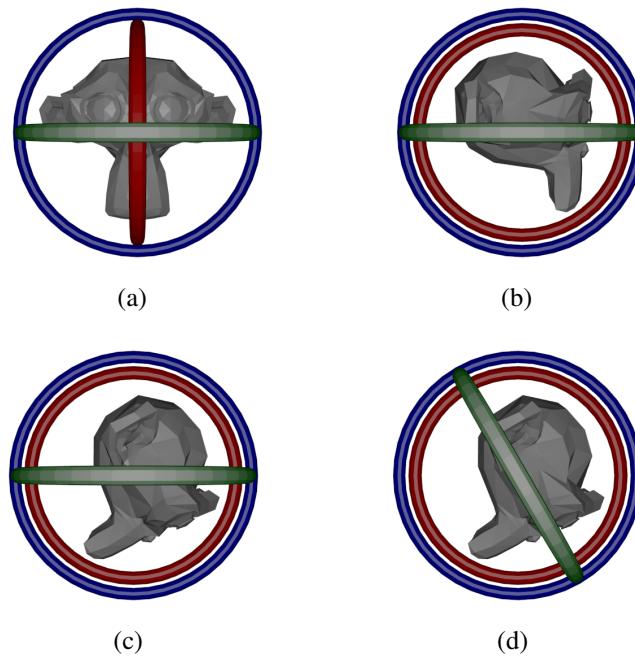


Figura 6.6: Los ejes rojo, verde y azul, son los ejes x , y y z . En (a) se ve la posición inicial. En (b) se puede ver el eje y rotado 90° . En (c) se ve la rotación según x de 60° respecto a la posición de (b). En (d) se ve la rotación según z de -60° respecto a la posición de (b).

Esto es una rotación entorno al vector $(0, 0, -1)$ de un ángulo $\alpha = \psi - \phi$

Esto se puede ver gráficamente en la Figura 6.7.2.3. Se realiza la rotación según y y se pueden ver que los ejes de x y z quedan alineados. Luego se realiza una rotación de 60° en torno a x y por otra parte también se hace otra igual pero de signo opuesto en torno a z . Se ve que la posición final, partiendo de los ejes alineados para una y otra rotación del modelo es la misma. Lo que cambia es la posición del eje y . Cuando se rota en torno a x , el eje y queda quieto porque está más arriba en la jerarquía de rotaciones para este caso particular. Cuando se rota en torno a z , el eje y se mueve ya que está por debajo de z en la jerarquía.

6.7.3. Cuaternios

Los cuaternios son una extensión a los números reales, son generados añadiendo las unidades imaginarias i , j y k . Se cumple que $i^2 = j^2 = k^2 = -1$. Un número cuaternion q se expresa como $q = a + bi + cj + dk$. También puede ser expresado como un escalar y un vector de 3 elementos (a, \mathbf{v}) . Una rotación alrededor del versor ω un ángulo θ se puede expresar como el número cuaternion unidad

$$q = \left(\cos\left(\frac{1}{2}\theta\right), \omega \sin\left(\frac{1}{2}\theta\right) \right)$$

Para rotar un punto 3D \mathbf{M} , se representa como un cuaternion $p = (0, \mathbf{M})$ y el punto p' rotado se calcula como

$$p' = qp\bar{q}.$$

El producto que se utiliza es el producto de cuaternios y $\bar{q} = (\cos\left(\frac{1}{2}\theta\right), -\omega \sin\left(\frac{1}{2}\theta\right))$ es el cuaternion conjugado de q .

Esta representación evita el problema del *gimbal lock* pero tiene como contra que tiene un mayor costo computacional.

CAPÍTULO 7

POSIT: *POS* with *ITerations*

7.1. Introducción

En este capítulo se explica el algoritmo utilizado para el cálculo de la pose a partir de una imagen capturada por la cámara. Como lo dice el nombre de algoritmo se utiliza una técnica llamada *POS* (*Pose from Orthography and Scaling*), esta técnica consiste en aproximar la pose de la cámara a partir de la proyección *SOP*(*Scaled Orthographic Projection*). Se comienza explicando la versión clásica de POSIT, en la cual se presentan las técnicas utilizadas dentro del algoritmo, entre ellas se explica en que consiste la proyección SOP. Luego se presenta el problema de trabajar con marcadores planos y se explica como se modifica el algoritmo para este caso. Se presenta también el algoritmo llamado SoftPOSIT, que sirve para obtener la pose en el caso en que no se conocen las correspondencias entre los puntos del modelo y los puntos detectados. También se presenta el una variación de POSIT que resuelve la estimación de una manera diferente a la versión clásica. Finalmente se presentan los resultados obtenidos de la comparación de las dos versiones de POSIT.

7.2. POSIT clásico

La primera versión de POSIT presentada por DeMenthon et al. en [?] resuelve el problema de estimar la pose de la cámara dados 4 o más puntos detectados en la imagen y sus correspondientes en el mundo real, con la condición de que estos puntos no sean coplanares. Si bien no es la versión final que se utilizó vale la pena ser explicada ya que ayuda a explicar los fundamentos del algoritmo utilizado.

7.2.1. Notación

En la Figura 7.1 se puede ver un modelo de cámara pinhole como el que se presentó en 6.2.1. O es el centro óptico y G es el plano imagen ubicado a una distancia focal f de O . x e y son los ejes que apuntan en las direcciones de las filas y las columnas del sensor de la cámara respectivamente. z es el eje que esta sobre el eje óptico de la cámara y apunta en sentido saliente. Los versores para estos ejes son \mathbf{i} , \mathbf{j} y \mathbf{k} respectivamente.

Se considera ahora un objeto con puntos característicos $M_0, M_1, \dots, M_i, \dots, M_n$, cuyo eje de coordenadas esta centrado en M_0 y está compuesto por los versores (M_0u, M_0v, M_0w) . Como el los ejes del mundo son arbitrarios se toma que los ejes del objeto son los ejes del mundo. La geometría del objeto se asume conocida, por lo tanto las coordenadas de los puntos característicos del objeto en el eje de coordenadas del mismo son conocidas. Por ejemplo (U_i, V_i, W_i) son las coordenadas del

punto M_i en el marco de referencia del objeto. Los puntos correspondientes a los puntos del objeto M_i en la imagen son conocidos y se identifican como m_i , (x_i, y_i) son las coordenadas de este punto en la imagen¹. Las coordenadas de los puntos M_i en el eje de coordenadas de la cámara, identificadas como (X_i, Y_i, Z_i) , son desconocidas ya que no se conoce la pose del objeto respecto a la cámara.

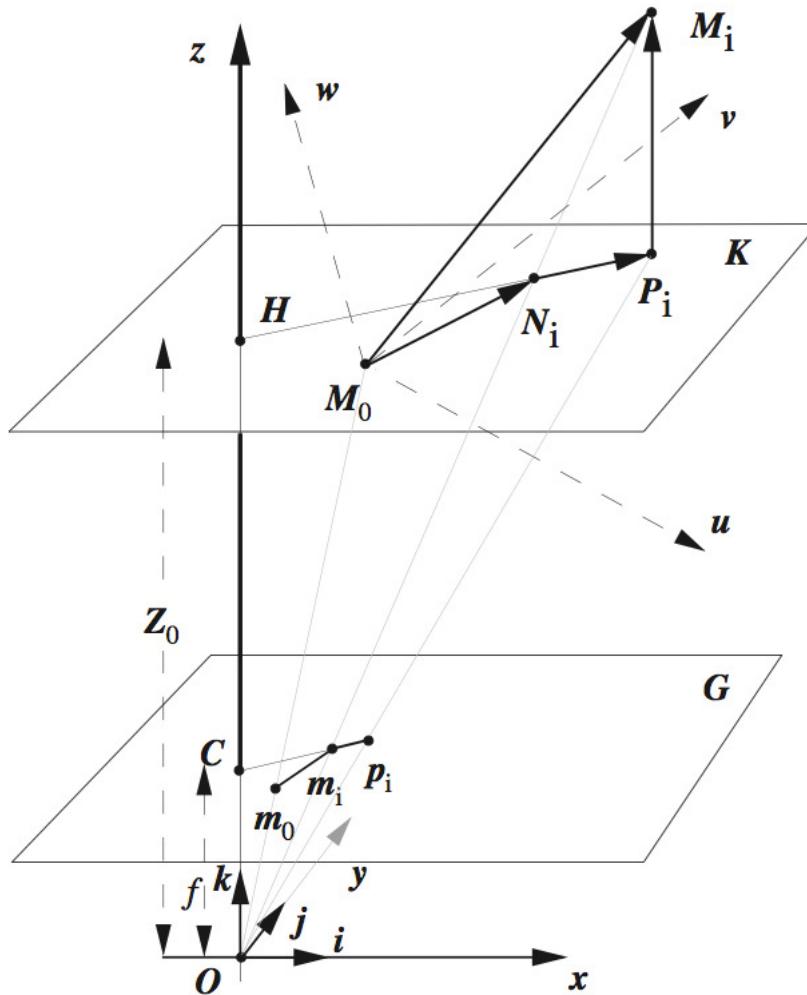


Figura 7.1: Proyección en perspectiva (m_i) y SOP (p_i) para un punto del modelo 3D M_i y un punto de referencia del modelo M_0 . Tomado de: [?].

Se busca computar la matriz de rotación y el vector de translación del objeto respecto a la cámara. Se recuerda que matriz de rotación se expresa como:

$$R = \begin{pmatrix} i_u & i_v & i_w \\ j_u & j_v & j_w \\ k_u & k_v & k_w \end{pmatrix}$$

Por lo tanto, para obtener la matriz de rotación sólo es necesario obtener los versores \mathbf{i} y \mathbf{j} , el versor \mathbf{k} se obtiene de realizar el producto vectorial $\mathbf{i} \times \mathbf{j}$. El vector de translación es el vector que va del centro del objeto M_0 a el centro del sistema de coordenadas de la cámara O . Por lo tanto las coordenadas del vector de translación son (X_0, Y_0, Z_0) . Si este punto M_0 es uno de los puntos visibles en la imagen, entonces el vector \mathbf{T} esta alineado con el vector Om_0 y es igual a $(Z_0/f)Om_0$. La pose queda determinada si se conocen \mathbf{i} , \mathbf{j} y Z_0 .

¹En realidad los puntos m_i no están dados, vienen de la etapa anterior de detección.

7.2.2. SOP: Scaled Orthographic Projection

La proyección ortogonal escalada(SOP) es una aproximación a la proyección perspectiva. En esta aproximación se supone que las profundidades Z_i de diferentes puntos M_i en el eje de coordenadas de la cámara no difieren mucho entre sí, y por lo tanto se asume que todos los puntos M_i tienen la misma profundidad que el punto M_0 . Esta suposición es razonable cuando la relación distancia cámara objeto - profundidad del objeto es grande.

Para un punto M_i la proyección perspectiva sobre el plano imagen estaría dada por:

$$x_i = fX_i/Z_i, \quad y_i = fY_i/Z_i,$$

mientras que la proyección SOP esta dada por:

$$x'_i = fX_i/Z_0, \quad y'_i = fY_i/Z_0.$$

De aquí en más las proyecciones SOP de los puntos M_i se identificarán como p_i , mientras que las proyecciones perspectivas, que son los puntos que se detectan en la imagen, se identifican como m_i . Al término $s = f/Z_0$ se lo conoce como el factor de escala de la SOP. Se puede ver que para el caso particular del punto M_0 la proyección perspectiva m_0 y la SOP p_0 coinciden.

En la Figura 7.1 se puede ver como se construye la SOP. Primero se realiza la proyección ortogonal de todos los puntos M_i sobre K , el plano paralelo al plano imagen que pasa por el punto M_0 . Las proyecciones de los puntos M_i sobre K se llaman P_i . El segundo paso consiste en hacer la proyección perspectiva de los puntos P_i sobre el plano imagen G para obtener finalmente los puntos p_i . En la figura también se puede ver que el tamaño del vector $m_0 p_i$ es s veces el tamaño de $M_0 P_i$. Teniendo esto en cuenta se pueden expresar las coordenadas de p_i como:

$$\begin{aligned} x'_i &= fX_0/Z_0 + f(X_i - X_0)/Z_0 = x_0 + s(X_i - X_0) \\ y'_i &= y_0 + s(Y_i - Y_0) \end{aligned}$$

7.2.3. Ecuaciones para calcular la proyección perspectiva

Como se mencionó anteriormente la pose queda determinada si se conocen los vectores \mathbf{i}, \mathbf{j} y la coordenada Z_0 del vector de traslación. La relación entre las coordenadas de los puntos M_i en el sistema de coordenadas del objeto y el sistema de coordenadas de la cámara es

$$\begin{pmatrix} X_i \\ Y_i \\ Z_i \end{pmatrix} = \begin{pmatrix} i_u & i_v & i_w \\ j_u & j_v & j_w \\ k_u & k_v & k_w \end{pmatrix} \begin{pmatrix} U_i \\ V_i \\ W_i \end{pmatrix} + \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix} \quad (7.1)$$

de esta expresión se tiene

$$x_i = f \frac{\mathbf{M}_0 \mathbf{M}_i \cdot \mathbf{i} + X_0}{\mathbf{M}_0 \mathbf{M}_i \cdot \mathbf{k} + Z_0}$$

y si saca Z_0 de factor común se tiene

$$x_i = \frac{\mathbf{M}_0 \mathbf{M}_i \cdot \frac{f}{Z_0} \mathbf{i} + x_0}{\mathbf{M}_0 \mathbf{M}_i \cdot \frac{f}{Z_0} \mathbf{k} + 1}$$

lo mismo se tiene para y_i .

Por lo tanto la condición necesaria para que la pose definida por \mathbf{i}, \mathbf{j} , x_0 , y_0 y Z_0 sea la pose exacta se puede expresar en las siguientes ecuaciones:

$$M_0 M_i \frac{f}{Z_0} \mathbf{i} = x_i(1 + \varepsilon_i) - x_0 \quad (7.2)$$

$$M_0 M_i \frac{f}{Z_0} \mathbf{j} = y_i(1 + \varepsilon_i) - y_0 \quad (7.3)$$

donde ε_i se define como

$$\varepsilon_i = \frac{1}{Z_0} M_0 M_i \cdot \mathbf{k} \quad (7.4)$$

Se puede ver que los términos $x_i(1 + \varepsilon_i)$ y $y_i(1 + \varepsilon_i)$ son las coordenadas (x'_i, y'_i) de la SOP, en el caso en que la pose esta determinada. En la expresión de ε_i en la Ecuación 10.7, el producto escalar con \mathbf{k} da la coordenada z de $M_0 M_i$, $Z_i - Z_0$, entonces se tiene que

$$(1 + \varepsilon_i) = \frac{Z_i - Z_0}{Z_0} + 1 = \frac{Z_i}{Z_0}$$

ademas se tiene la proyección perspectiva $x_i = f X_i / Z_i$, combinando las dos expresiones se tiene

$$x_i(1 + \varepsilon_i) = f \frac{X_i}{Z_i} \frac{Z_i}{Z_0} = f \frac{X_i}{Z_0}$$

que es la coordenada x'_i del punto p_i .

7.2.4. Algoritmo

Las Ecuaciones 10.2 y 10.3 se puede reescribir como:

$$M_0 M_i \mathbf{I} = x_i(1 + \varepsilon_i) - x_0 \quad (7.5)$$

$$M_0 M_i \mathbf{J} = y_i(1 + \varepsilon_i) - y_0 \quad (7.6)$$

en donde

$$\mathbf{I} = \frac{f}{Z_0} \mathbf{i} = s \cdot \mathbf{i}, \quad \mathbf{J} = \frac{f}{Z_0} \mathbf{j} = s \cdot \mathbf{j} \quad (7.7)$$

Si se conociera el valor de ε_i , las Ecuaciones 7.5 y 7.6 representan un sistema de ecuaciones en que las incógnitas son los vectores \mathbf{I} y \mathbf{J} . Una vez obtenidos estos vectores se pueden obtener los versores \mathbf{i} y \mathbf{j} normalizando, y Z_0 se obtiene de la norma de cualquiera de los vectores \mathbf{I} o \mathbf{J} . A esta parte del algoritmo se le llama *POS* (*Pose from Orthography and Scaling*), ya que estima la pose a partir de las proyecciones SOP de los puntos M_i .

Si se conocieran los valores exactos de los ε_i la pose obtenida de resolver el sistema de ecuaciones sería la pose exacta del objeto, como no se conocen los valores exactos de ε_i se utiliza un método iterativo que converge a la solución buscada. En la primera iteración se le toma $\varepsilon_i = 0$. La ecuación para un punto cualquiera está dada por:

$$\begin{aligned} M_0 M_i \cdot \mathbf{I} &= x'_i - x_0 \\ M_0 M_j \cdot \mathbf{J} &= y'_i - y_0 \end{aligned} \quad (7.8)$$

Si se escribe la Ecuación 7.8 para los n puntos del modelo, se tiene un sistema de n ecuaciones con \mathbf{I} y \mathbf{J} como incógnitas

$$\begin{aligned} \mathbf{A}\mathbf{I} &= x' - x_0 \\ \mathbf{A}\mathbf{J} &= y' - y_0 \end{aligned} \quad (7.9)$$

A es una matriz $n \times 3$ con las coordenadas de los puntos del modelo M_i en el marco de coordenadas del objeto. Si se tienen mas de 4 puntos y no son coplanares, la matriz **A** es de rango 3, y las soluciones al sistema están dadas por

$$\begin{aligned}\mathbf{I} &= \mathbf{B} \left(\mathbf{x}' - \mathbf{x}_0 \right) \\ \mathbf{J} &= \mathbf{B} \left(\mathbf{y}' - \mathbf{y}_0 \right)\end{aligned}\tag{7.10}$$

donde **B** es la pseudo inversa de la matriz **A**. Se debe notar que la matriz **B** depende únicamente de la geometría del modelo que se asume conocida, por lo tanto solo es necesario calcularla una sola vez.

Una vez obtenidos **I** y **J** se calculan s y los versores **i**, **j** y **k**

$$s = (|\mathbf{I}| |\mathbf{J}|)^{1/2}\tag{7.11a}$$

$$i = \frac{\mathbf{I}}{s}\tag{7.11b}$$

$$j = \frac{\mathbf{J}}{s}\tag{7.11c}$$

$$k = i \times j\tag{7.11d}$$

El vector traslación del centro del objeto al centro de la cámara es el vector OM_0

$$OM_0 = \frac{Z_0}{f} Om_0 = \frac{Om_0}{s}\tag{7.12}$$

El vector Om_0 es conocido ya que se conocen las coordenadas de los puntos m_i , en particular m_0 .

Una vez que se calcularon **i**, **j**, **k** y **T** se calculan los valores actualizados de ε_i según la Ecuación 10.7. Si la variación de los ε_i es mayor a un determinado umbral, se repite el procedimiento actualizando las proyecciones SOP en la Ecuación 7.9, si es menor al umbral se deja de iterar y se guarda la pose calculada.

7.2.5. POSIT para puntos coplanares

Como se mencionó anteriormente, el algoritmo POSIT no funciona en el caso en que los puntos del modelo pertenecen a un mismo plano. Como los marcadores utilizados son planos, se buscó una versión de POSIT que resuelve el problema de la estimación de pose para este caso. El algoritmo fue escrito por DeMenthon et al. en [?].

Para entender cual es el problema de trabajar con puntos coplanares se explica la situación desde un punto de vista geométrico. Como se vió anteriormente,

$$M_0 M_i \cdot \mathbf{I} = \mathbf{x}'_i - \mathbf{x}_0.$$

Esto quiere decir que si se toma que la base de **I** en M_0 , la punta del vector **I** se proyecta sobre el vector $M_0 M_i$ en un punto H_{x_i} , entonces todas las posibles puntas del vector **I** se encuentran en el plano perpendicular a $M_0 M_i$ que pasa por el punto H_{x_i} . Si se tuvieran 4 puntos no coplanares M_0, M_1, M_2 y M_3 , el vector **I** quedaría determinado. La base de **I** estaría en M_0 y la punta estaría en la intersección de los planos perpendiculares a $M_0 M_1, M_0 M_2$ y $M_0 M_3$ por los puntos H_{x_1}, H_{x_2} y H_{x_3} respectivamente. Para este caso el sistema definido en 7.9 es de rango 3.

Si los puntos son coplanares, los vectores $M_0 M_1, M_0 M_2$ y $M_0 M_3$ son todos coplanares y los planos perpendiculares que pasan por los puntos H_{x_1}, H_{x_2} y H_{x_3} , se intersectan todos en una línea o

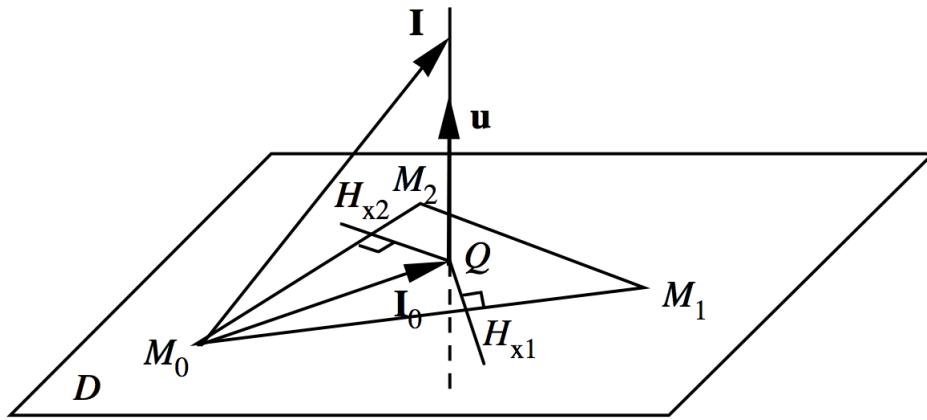


Figura 7.2: Configuración de puntos coplanares pertenecientes al plano D . Los planos perpendiculares que pasan por los puntos H_{x1} y H_{x2} se intersectan en un recta que pasa por el punto Q . Se puede ver que si hubiera un 4^{to} punto, el plano perpendicular correspondiente haría aparecer 2 rectas paralelas. Tomado de: [?].

en dos líneas paralelas por lo tanto hay infinitas soluciones para el vector \mathbf{I} . En este caso el sistema de ecuaciones en 7.9 queda de rango 2. El vector solución que se obtiene al realizar la pseudo inversa de \mathbf{A} es el que está a menor distancia de los planos, en la Figura 7.2 es el vector \mathbf{I}_0 . Esta la solución no es la solución al problema de los vectores de rotación, las soluciones se pueden expresar como

$$\begin{aligned}\mathbf{I} &= \mathbf{I}_0 + \lambda \mathbf{u} \\ \mathbf{J} &= \mathbf{J}_0 + \mu \mathbf{u}\end{aligned}\tag{7.13}$$

donde \mathbf{u} es un versor perpendicular al plano de los puntos, \mathbf{J}_0 se calcula de manera análoga a \mathbf{I}_0 y λ y μ son las coordenadas de \mathbf{I} y \mathbf{J} según el versor \mathbf{u} . Para encontrar las soluciones hay que calcular el versor \mathbf{u} y los valores de λ y μ .

Como el vector \mathbf{u} es perpendicular al plano de los puntos característicos se cumple $M_0M_i \cdot \mathbf{u} = 0$, se puede hallar entonces como la base del núcleo de la matriz \mathbf{A} . En la práctica este vector se halla a partir de la descomposición SVD de la matriz \mathbf{A} . La descomposición en valores singulares de la matriz \mathbf{A} queda:

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T\tag{7.14}$$

donde $\mathbf{U} \in \mathbb{R}^{n \times n}$ es ortogonal, $\Sigma \in \mathbb{R}^{n \times 3}$ es diagonal, con los valores singulares en la diagonal y $\mathbf{V} \in \mathbb{R}^{3 \times 3}$ es ortogonal. Como la matriz \mathbf{A} es de rango 2, los dos primeros vectores columna de la matriz \mathbf{V} corresponden a la base de todos los puntos que pertenecen al plano del modelo, mientras que el último vector de \mathbf{V} es la base del núcleo de \mathbf{A} , o sea el vector \mathbf{u} . El cálculo \mathbf{u} se realiza junto al cálculo de la matriz \mathbf{B} , ya que para ambos es necesario hacer la descomposición SVD de \mathbf{A} .

Para calcular los valores de λ y μ se utilizan las condiciones de que \mathbf{I} y \mathbf{J} tienen que ser perpendiculares entre sí y del mismo largo. Como tienen que ser perpendiculares se tiene que

$$\mathbf{I} \cdot \mathbf{J} = (\mathbf{I}_0 + \lambda \mathbf{u}) \cdot (\mathbf{J}_0 + \mu \mathbf{u}) = 0$$

entonces se tiene que

$$\lambda \mu = -\mathbf{I}_0 \cdot \mathbf{J}_0\tag{7.15}$$

Como tienen que ser del mismo largo se tiene que

$$(\mathbf{I}_0 + \lambda \mathbf{u}) \cdot (\mathbf{I}_0 + \lambda \mathbf{u}) = (\mathbf{J}_0 + \mu \mathbf{u}) \cdot (\mathbf{J}_0 + \mu \mathbf{u}) \Leftrightarrow \lambda^2 - \mu^2 = \mathbf{J}_0^2 - \mathbf{I}_0^2\tag{7.16}$$

Se define el número complejo $C = \lambda + i\mu$, si se eleva al cuadrado queda $C^2 = \lambda^2 - \mu^2 + i\lambda\mu$. Utilizando 7.15 y 7.16 se llega a que

$$C^2 = \mathbf{J}_0^2 - \mathbf{I}_0^2 - 2i\mathbf{I}_0 \cdot \mathbf{J}_0 \quad (7.17)$$

por lo que λ y μ pueden calcularse como las partes real e imaginaria del complejo C^2 . Para hallar la raíces de C^2 , se expresa en forma polar:

$$\begin{aligned} C^2 &= [R, \Theta], \text{ donde} \\ R &= \left((\mathbf{J}_0^2 - \mathbf{I}_0^2)^2 + 4(\mathbf{I}_0 \cdot \mathbf{J}_0)^2 \right)^{1/2} \\ \Theta &= \arctan \left(\frac{-2\mathbf{I}_0 \cdot \mathbf{J}_0}{\mathbf{J}_0^2 - \mathbf{I}_0^2} \right), \text{ si } \mathbf{J}_0^2 - \mathbf{I}_0^2 > 0, \text{ y} \\ \Theta &= \arctan \left(\frac{-2\mathbf{I}_0 \cdot \mathbf{J}_0}{\mathbf{J}_0^2 - \mathbf{I}_0^2} \right) + \pi, \text{ si } \mathbf{J}_0^2 - \mathbf{I}_0^2 < 0 \\ \text{si } \mathbf{J}_0^2 - \mathbf{I}_0^2 = 0 \text{ se toma } \Theta &= -sg(\mathbf{I}_0 \cdot \mathbf{J}_0) \frac{\pi}{2}, \text{ y } R = |2\mathbf{I}_0 \cdot \mathbf{J}_0| \end{aligned}$$

Se obtienen 2 raíces, $C = [\rho, \theta]$, y $C = [\rho, \theta + \pi]$, donde

$$\rho = \sqrt{R}, \text{ y } \theta = \frac{\Theta}{2}$$

como se mencionó anteriormente λ y μ son las partes real e imaginaria de C , por lo tanto

$$\lambda_1 = \rho \cos \theta, \quad \mu_1 = \rho \sin \theta \quad (7.18a)$$

$$\lambda_2 = -\rho \cos \theta, \quad \mu_2 = -\rho \sin \theta \quad (7.18b)$$

Esto quiere decir que se obtienen dos soluciones para \mathbf{I} y \mathbf{J}

$$\mathbf{I}_1 = \mathbf{I}_0 + \rho \cos \theta \mathbf{u}, \quad \mathbf{J}_1 = \mathbf{J}_0 + \rho \sin \theta \mathbf{u} \quad (7.19a)$$

$$\mathbf{I}_2 = \mathbf{I}_0 - \rho \cos \theta \mathbf{u}, \quad \mathbf{J}_2 = \mathbf{J}_0 - \rho \sin \theta \mathbf{u} \quad (7.19b)$$

Como el vector \mathbf{u} es perpendicular al plano del objeto, la solución encontrada en 7.19a es simétrica a 7.19b. Desde el punto de vista de la cámara, se puede ver que las dos posibles soluciones son aquellas que tienen la misma proyección SOP, este comportamiento se puede ver en la Figura 7.3. Esto es equivalente a decir que para una misma proyección SOP hay dos posibles poses que verifican las ecuaciones en 7.9.

Por lo tanto se toman las soluciones $(\mathbf{I}_1, \mathbf{J}_1)$ y $(\mathbf{I}_2, \mathbf{J}_2)$ y se calculan las poses. Como las dos poses son simétricas respecto a un plano paralelo al plano imagen, puede pasar que una pose dé una solución en la que los puntos del objeto queden ubicados detrás de la cámara. Por lo tanto previo a dar las dos soluciones como válidas hay que verificar esto.

En el caso en que las dos soluciones sean válidas para todas las iteraciones, el número de poses posibles sería 2^n a lo largo de n iteraciones. En la práctica se manejan menos soluciones posibles. Se diferencian dos casos:

- Si se tiene que sólo una de las dos primeras poses calculadas es válida, en las siguientes iteraciones se da el mismo comportamiento, por lo que hay solo un camino a seguir. Figura 7.4(a).
- Si se tiene que las dos primeras poses calculadas son válidas, se abren dos posibles ramas. En la segunda iteración cada rama da lugar a dos nuevas poses, pero en este caso se toma la pose que da menor error de reproyección. Figura 7.4(b).

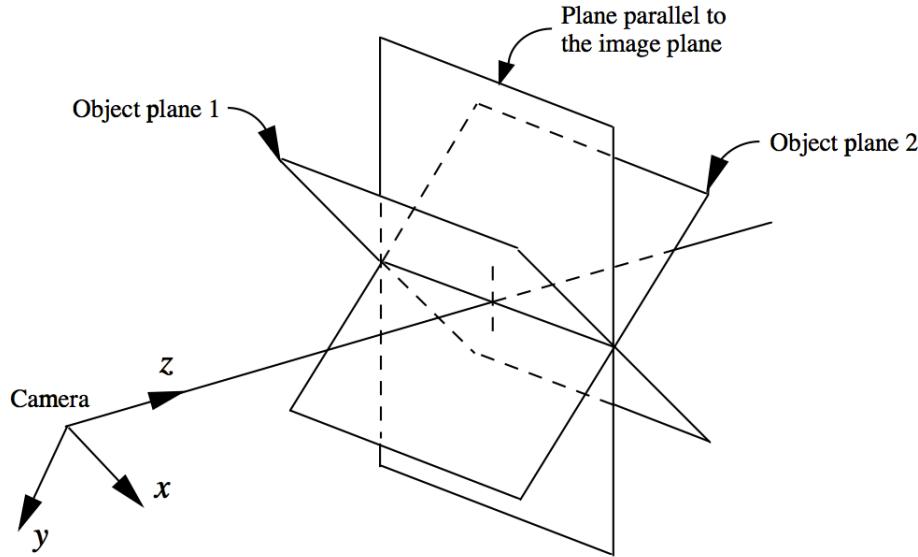


Figura 7.3: Dos objetos dando la misma proyección SOP. Fuente: [?].

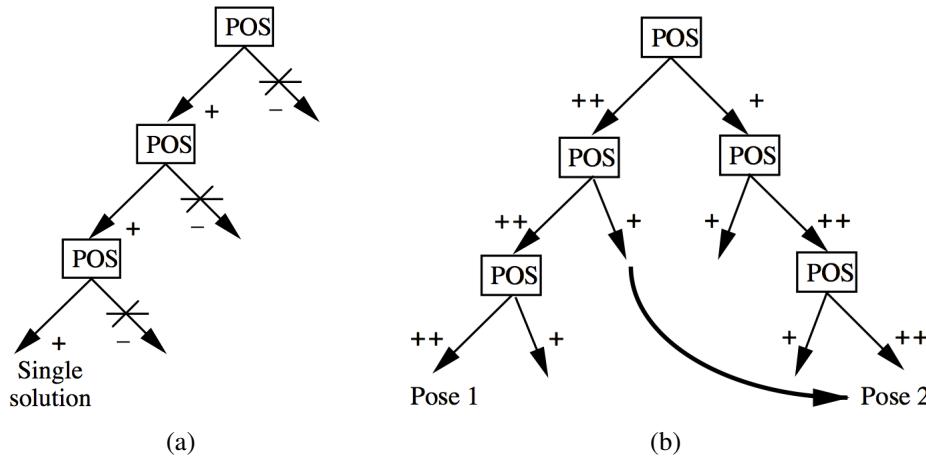


Figura 7.4: (a):Caso en el que solo una pose de las dos iniciales es coherente, también en las siguientes iteraciones solo una de las dos poses es posible, se tiene un única solución. (b): Caso en el que en cada paso hay dos posibilidades, se opta por la mejor pose(++) mejor pose, + peor pose) en cada rama. Tomado de: [?].

7.3. SoftPOSIT

Hasta aquí se vió el algoritmo POSIT que permite obtener la pose de un modelo respecto a la cámara para el caso en que se tienen correspondencias entre puntos del modelo y puntos característicos en la imagen. Como se vió en el capítulo 3 obtener correspondencias entre puntos detectados en una imagen y el modelo real puede ser complicado. Por este motivo se estudió el algoritmo SoftPOSIT desarrollado por DeMenthon et al. en [4]. Este algoritmo recibe como entrada los puntos del modelo 3D y una lista de puntos detectados en la imagen para los cuales no se sabe como se relacionan con los puntos del modelo. Utiliza un método llamado *softassign* para resolver las correspondencias y luego que tiene las correspondencias utiliza una versión modificada de POSIT.

7.3.1. Modern POSIT

Como se mencionó anteriormente, SoftPOSIT utiliza una versión modificada de POSIT llamada Modern POSIT. POSIT clásico requiere que se conozca cual es el punto de referencia en el modelo y en la imagen, ya que de estos datos se calcula el vector de traslación. Para el caso de SoftPOSIT no es posible saber de antemano cual es el punto de referencia del modelo ya que no se tienen las correspondencias. Modern POSIT calcula la pose, sabiendo las correspondencias, pero sin utilizar ningún punto en particular como referencia. Además la pose se calcula minimizando una función que mide la distancia entre la proyección SOP y los puntos estimados en cada iteración.

El punto M_0 origen del sistema de coordenadas del objeto no es conocido, por lo tanto tampoco se conoce su correspondiente m_0 en el plano imagen. En la Ecuación 7.8 que se presentó en la sección 7.2.4 se conocían las coordenadas del punto m_0 , por lo que las incógnitas de esta ecuación eran solamente los vectores \mathbf{i}, \mathbf{j} .

$$\begin{aligned} M_0M_i \cdot \mathbf{I} &= x'_i - x_0 \\ M_0M_j \cdot \mathbf{J} &= y'_i - y_0 \end{aligned}$$

En este caso no se conocen las coordenadas de m_0 , por lo que también hace falta calcularlas para obtener el vector de traslación. Sabiendo que

$$\begin{aligned} X_0 &= x_0/s \\ Y_0 &= y_0/s \end{aligned}$$

se puede reescribir la Ecuación 7.8 como

$$\begin{aligned} x'_i &= M_0M_i \cdot s\mathbf{i} + sX_0 \\ y'_i &= M_0M_j \cdot s\mathbf{j} + sY_0 \end{aligned} \tag{7.20}$$

El sistema a resolver queda

$$\mathbf{A} \cdot \begin{bmatrix} \mathbf{I} & \mathbf{J} \\ sX_0 & sY_0 \end{bmatrix} = \begin{bmatrix} x' & y' \end{bmatrix} \tag{7.21}$$

donde la matriz \mathbf{A} son los puntos del modelo 3D en coordenadas homogéneas. Este sistema se puede resolver, utilizando mínimos cuadrados, como se vió en la sección 7.2.4.

Sin embargo se propone un método que busca minimizar la distancia al cuadrado entre las proyecciones SOP de los puntos M_i y las proyecciones SOP calculadas en cada iteración. En la Figura 7.3.1 se puede ver geométricamente cual es la distancia que se busca minimizar.

En el término de la derecha de la Ecuación 7.20 se tiene la proyección SOP de M_i , p_i . Las coordenadas de este punto son

$$p_i = s(M_0M_i \cdot \mathbf{i} + X_0, M_0M_i \cdot \mathbf{j} + Y_0).$$

Por otro lado, en el término de la izquierda de 7.20 se tienen las coordenadas del punto p'_i

$$p'_i = (1 + \varepsilon_i)(x_i, y_i).$$

que es la proyección SOP de la intersección de la línea de vista del punto m_i con el plano G'' , esto está demostrado en [5]. La pose encontrada es correcta cuando ambos lados de la Ecuación 7.20 son iguales. Por lo tanto la ecuación que se busca minimizar es la siguiente:

$$E = \sum_i \left((\mathbf{Q}_1 \cdot M_0M_i - (1 + \varepsilon_i)x_i)^2 + (\mathbf{Q}_2 \cdot M_0M_i - (1 + \varepsilon_i)y_i)^2 \right) \tag{7.22}$$

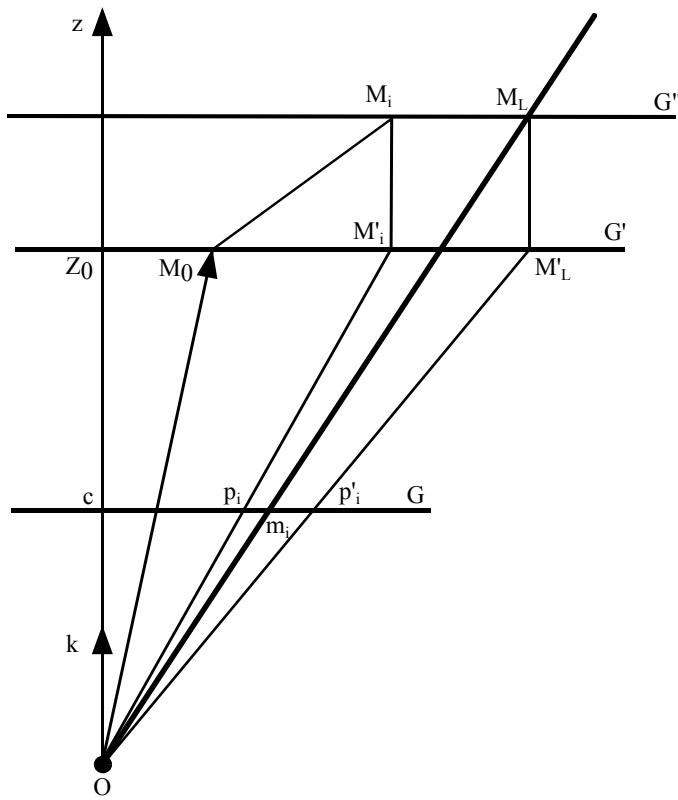


Figura 7.5: Interpretación geométrica de POSIT. El punto p_i es la proyección SOP de M_i que es el término de la derecha de la Ecuación 7.20. El punto p'_i es la proyección SOP de M_L , ubicado en la línea de vista de m_i , corresponde al término de la izquierda de la Ecuación 7.20. Para que la ecuación se satisfaga se tiene que cumplir que p_i y p'_i sean iguales. Fuente: [5].

donde

$$\begin{aligned}\mathbf{Q}_1 &= s(i, X_0) \\ \mathbf{Q}_2 &= s(j, Y_0)\end{aligned}\tag{7.23}$$

y M_0M_i se toma en coordenadas homogéneas.

Los vectores \mathbf{Q}_1 y \mathbf{Q}_2 son aquellos que minimizan el valor E , por lo tanto se despejan de derivar la expresión de E e igualarla a cero. La expresión para calcular \mathbf{Q}_1 y \mathbf{Q}_2 queda:

$$\mathbf{Q}_1 = \left(\sum_i M_0 M_i^T M_0 M_i \right)^{-1} \left(\sum_i (1 + \varepsilon_i) x_i M_0 M_i \right) \tag{7.24}$$

$$\mathbf{Q}_2 = \left(\sum_i M_0 M_i^T M_0 M_i \right)^{-1} \left(\sum_i (1 + \varepsilon_i) y_i M_0 M_i \right) \tag{7.25}$$

La matriz $L = (\sum_i M_0 M_i^T M_0 M_i)$ es una matriz 4×4 y como sólo depende de los puntos del modelo puede ser calculada previamente.

Para calcular la pose se procede como sigue:

- (1) Se calculan los vectores \mathbf{Q}_1 y \mathbf{Q}_2 asumiendo que se conocen los valores de ε_i , para el paso inicial se supone que $\varepsilon_i = 0$.
- (2) Con los vectores \mathbf{Q}_1 y \mathbf{Q}_2 calculados se calculan los ε_i corregidos.

Cuando E es menor a determinado umbral el algoritmo se detiene y se obtiene la pose calculada.

7.3.2. Cálculo de pose sin correspondencias

Se tienen N puntos detectados en la imagen y M puntos en el modelo. Cuando no se conocen las correspondencias cada punto detectado p_j es candidato a corresponderse con cualquier punto del modelo M_i . La distancia que se busca minimizar es

$$d_{ji}^2 = (\mathbf{Q}_1 \cdot M_i M_0 - (1 + \varepsilon_i) x_j)^2 + (\mathbf{Q}_2 \cdot M_i M_0 - (1 + \varepsilon_i) y_j)^2$$

Se puede ver que para cada punto de modelo hay N candidatos, la distancia d_{ji}^2 da una idea de que tan cerca esta de ser el correspondiente. Para resolver el problema de estimar la pose y las correspondencias simultáneamente se busca minimizar la siguiente función:

$$\begin{aligned} E &= \sum_{j=1}^N \sum_{i=1}^M m_{ji} (d_{ji}^2 - \alpha) \\ &= \sum_{j=1}^N \sum_{i=1}^M m_{ji} \left((\mathbf{Q}_1 \cdot M_0 M_i - (1 + \varepsilon_i) x_j)^2 + (\mathbf{Q}_2 \cdot M_0 M_i - (1 + \varepsilon_i) y_j)^2 - \alpha \right) \end{aligned} \quad (7.26)$$

donde m_{ji} son pesos para cada una de las distancias d_{ji}^2 . Los pesos m_{ji} forman lo que se llama matriz de asignación, en esta matriz se puede ver el grado de correspondencia de cualquier punto detectado con cualquier punto del modelo. El valor α es la tolerancia que se le da a la medida de distancia.

Las expresiones para los vectores \mathbf{Q}_1 y \mathbf{Q}_2 se modifican

$$\mathbf{Q}_1 = \left(\sum_{i=1}^M m'_i M_0 M_i^T M_0 M_i \right)^{-1} \left(\sum_{j=1}^N \sum_{i=1}^M m_{ji} (1 + \varepsilon_i) x_j M_0 M_i \right) \quad (7.27)$$

$$\mathbf{Q}_2 = \left(\sum_{i=1}^M m'_i M_0 M_i^T M_0 M_i \right)^{-1} \left(\sum_{j=1}^N \sum_{i=1}^M m_{ji} (1 + \varepsilon_i) y_j M_0 M_i \right) \quad (7.28)$$

donde $m'_i = \sum_{j=1}^N m_{ji}$. El término $L = \sum_{i=1}^M m'_i M_0 M_i^T M_0 M_i$ es una matriz 4×4 , para este caso L no se puede calcular previamente porque la matriz de asignación cambia en cada iteración.

Para minimizar E se procede como sigue:

- (1) Se calculan las variables de la matriz de asignación asumiendo todo lo demás conocido.
- (2) Se calculan los vectores \mathbf{Q}_1 y \mathbf{Q}_2 asumiendo que se conocen los valores de ε_i . Para el paso inicial se supone que $\varepsilon_i = 0$.
- (3) Con los vectores \mathbf{Q}_1 y \mathbf{Q}_2 calculados se calculan los ε_i corregidos.

Esto se repite hasta que la pose converge.

7.3.3. Matriz de asignación

Se busca tener una matriz m que indique las correspondencias entre los N puntos detectados y los M puntos en del modelo y además minimice E . La matriz de asignación tiene las siguientes características:

- Tiene $N+1$ filas y $M+1$ columnas.
- $m_{ji} \in [0, 1]$. Si $m_{ji} = 1$ quiere decir que el punto detectado p_j se corresponde con el punto del modelo M_i .

- La fila $N+1$ y la columna $M+1$ se utilizan para ver si alguna correspondencia en esa fila o columna. Por ejemplo si el elemento j de la columna $M+1$ es 1, significa que el punto detectado p_j no se corresponde con ningún punto del modelo.
- La suma de los elementos a lo largo de cualquier fila o columna es siempre 1.

Para obtener una matriz m que cumpla con las características mencionadas se utiliza una técnica llamada *softassign*. Se comienza con una matriz m^0 en la que los elementos están dados por

$$m_{ji}^0 = e^{-\beta(d_{ji}^2 - \alpha)}$$

en donde β es una constante muy pequeña y la fila $N+1$ y la columna $M+1$ son inicializadas con constantes pequeñas. Luego se itera utilizando los siguientes pasos hasta obtener la matriz m .

- (1) Se normaliza cada fila y columna por la suma de los elementos de esa fila o columna respectivamente hasta que $\|m^i - m^{i-1}\|$ sea pequeño. La matriz resultante cumple que todas las filas y columnas suman 1.
- (2) Se incrementa el valor de β a medida que se itera. A medida que se agranda β cada fila y columna de m^0 es renormalizada, los términos m_{ji}^0 correspondientes a las d_{ji}^2 convergen a 1, mientras que los demás convergen a 0.

Al final del algoritmo se observa que la matriz m está muy cerca de ser una matriz binaria indicando las correspondencias.

7.3.4. Implementación

Durante la investigación de este algoritmo se desarrollaron versiones de POSIT moderno y Soft-POSIT en C. Ambas implementaciones son autocontenido, no se necesita ninguna librería adicional para poderlas usar. Además todas las funciones están incluidas en un solo archivo. Esto es de gran valor ya que no se encontró en la web una versión de estos algoritmos que fuera del tipo *plug and play* como lo son estas.

7.4. POSIT moderno para puntos coplanares

La implementación que se usó en la aplicación es el POSIT moderno adaptado para trabajar con puntos coplanares. Inicialmente se quiso desarrollar una versión de SoftPOSIT que trabajara con puntos coplanares, para ello previamente se desarrolló POSIT moderno coplanar a modo de prueba.

Como se vió en la sección 7.2.5 cuando los puntos son coplanares, al resolver el sistema 7.9 se obtienen las proyecciones de los vectores **i** y **j** sobre el plano del objeto. Se utilizó el enfoque de POSIT moderno para hallar las proyecciones de **i** y **j** sobre el plano del modelo, así como los componentes en x e y del vector de translación. Luego aplicando lo visto en POSIT para puntos coplanares se terminó de calcular la pose.

Se definen los puntos $M_0M_i^*$ como los puntos M_0M_i sin la coordenada z , ya que la coordenada z es función de x e y . A su vez se definen los vectores \mathbf{Q}_1^* y \mathbf{Q}_2^* como los vectores \mathbf{Q}_1 y \mathbf{Q}_2 sin la componente según el eje w en el sistema de coordenadas del modelo. Teniendo esto en cuenta se tiene

$$E^* = \sum_i \left((\mathbf{Q}_1^* \cdot M_0M_i^* - (1 + \varepsilon_i)x_i)^2 + (\mathbf{Q}_2^* \cdot M_0M_i^* - (1 + \varepsilon_i)y_i)^2 \right)$$

Los vectores \mathbf{Q}_1^* y \mathbf{Q}_2^* se calculan de

$$\mathbf{Q}_1^* = \left(\sum_{i=1}^M m_i' M_0 M_i^{*T} M_0 M_i^* \right)^{-1} \left(\sum_{j=1}^N \sum_{i=1}^M m_{ji} (1 + \varepsilon_i) x_j M_0 M_i^* \right)$$

$$\mathbf{Q}_2^* = \left(\sum_{i=1}^M m_i' M_0 M_i^{*T} M_0 M_i^* \right)^{-1} \left(\sum_{j=1}^N \sum_{i=1}^M m_{ji} (1 + \varepsilon_i) y_j M_0 M_i^* \right)$$

En este caso el término $L = \sum_{i=1}^M m_i' M_0 M_i^{*T} M_0 M_i^*$ es una matriz 3×3 . Una vez que se tienen los vectores \mathbf{Q}_1^* y \mathbf{Q}_2^* se procede como se vió en la sección 7.2.5

Esta implementación de POSIT para puntos coplanares dió mejores resultados que la versión obtenida de [?]. Es una variante de POSIT coplanar que no se encuentra en la bibliografía, permite obtener la pose minimizando una función de costo a diferencia de POSIT clásico que usa mínimos cuadrados.

Además desde el punto de vista de programación se mejoró en la interfaz respecto a la versión clásica. En la versión clásica se tiene varios archivos con las diferentes funciones que se necesitan, puede llegar a ser difícil entender por completo la arquitectura del algoritmo y el código esta comentado en francés. En cambio en la versión implementada para este proyecto se busco tener un solo archivo con todas las funciones y mejorar la arquitectura respecto a la versión anterior.

7.5. Resultados

Se realizó un comparación entre la implementación en C de POSIT clásico para puntos coplanares, obtenida de [?] , y una versión desarrollada para esta aplicación de POSIT moderno para puntos coplanares.

Se utilizaron imágenes de prueba obtenidas con el *iPad* 2 e imágenes sintéticas. Para ambos grupos de imágenes se midió el error de proyección obtenido entre los puntos del modelo y los puntos detectados. Para las imágenes sintéticas, como se cuenta con la información de la pose, se midió el error obtenido en cada ángulo y en la traslación.

	Modern POSIT	Varianza	Classic POSIT	Varianza
Caso1	3.6136	0.8104	4.5979	1.1392
Caso2	0.8449	0.4153	0.9275	0.4415
Caso3	-	-	-	-
Caso4	1.5894	0.2600	1.1081	0.1696
Caso5	-	-	-	-
Caso6	-	-	-	-
Caso7	0.6742	0.1468	0.5416	0.1022
Caso8	-	-	-	-
Caso9	-	-	-	-

Tabla 7.1: Error de proyección de imágenes de pruebas.

Para el caso de las imágenes de prueba del *iPad* se eligieron 9 posiciones y en cada posición se sacaron 50 fotos. Con estas 450 fotos se obtuvo la estadística del funcionamiento de los algoritmos. En la Figura 7.5 se puede ver una de las imágenes utilizadas para cada posición. También se utilizaron imágenes sintéticas en posiciones similares a las de las imágenes de prueba, se utilizaron 9 casos con 50 fotos por caso. Se partió de una posición base y se varió la posición muy poco,

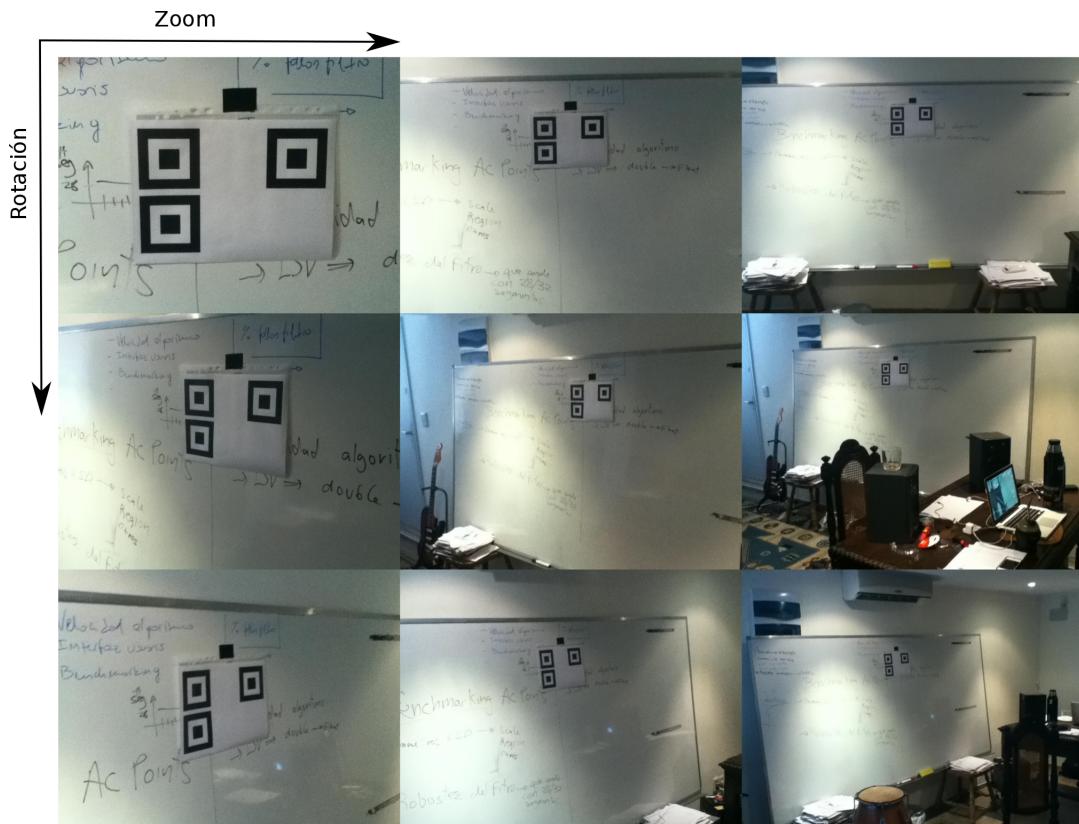


Figura 7.6: Posiciones que se utilizaron para las imágenes de prueba.

intentando simular el movimiento que se tuvo al sacar las fotos con el *iPad*. En total se probaron 900 imágenes.

A las imágenes se les aplica todo el proceso, se realiza la detección y filtrado de segmentos, se calculan las correspondencias y luego se estima la pose. Para cada imagen se calcula el error de proyección de cada punto, luego se promedian obteniendo una sola medida de error por imagen, esta medida es a su vez promediada con los errores obtenidos de la imágenes para un mismo caso. En las tablas hay valores que no pudieron ser calculados debido a que el filtro de segmentos no pudo detectar todos los segmentos. Estos comportamientos fueron discutidos en 3. En general se puede ver que el error de proyección y la varianza son un poco menores para el caso de modern POSIT.

Para otro grupo de imágenes sintéticas se comparó la pose original con la pose obtenida luego

	Modern POSIT	Varianza	Classic POSIT	Varianza
Caso1	4.2712	0.3192	5.7352	0.4525
Caso2	1.0831	0.0375	1.0889	0.0358
Caso3	-	-	-	-
Caso4	0.7975	0.0169	0.9778	0.0185
Caso5	-	-	-	-
Caso6	-	-	-	-
Caso7	0.3796	0.0121	0.4761	0.0077
Caso8	-	-	-	-
Caso9	-	-	-	-

Tabla 7.2: Error de proyección en imágenes sintéticas

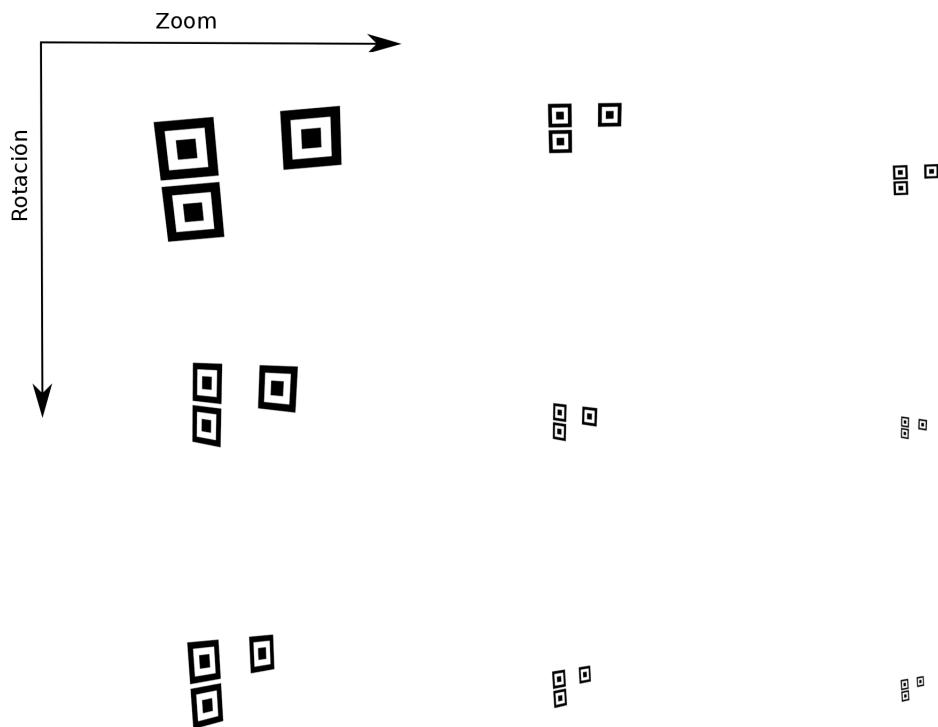


Figura 7.7: Posiciones que se utilizaron para las imágenes sintéticas

de aplicar el procesamiento, se relevó el desempeño de los algoritmos para rotaciones según los tres ejes. En general se vió que la implementación de POSIT moderno dio mejores resultados. En la Figura 7.5 se pueden algunos de los resultados obtenidos de la comparación de desempeño de los dos algoritmos implementados. En la columna de la izquierda se muestran los resultados de POSIT moderno y en la de la derecha los resultados para POSIT clásico.

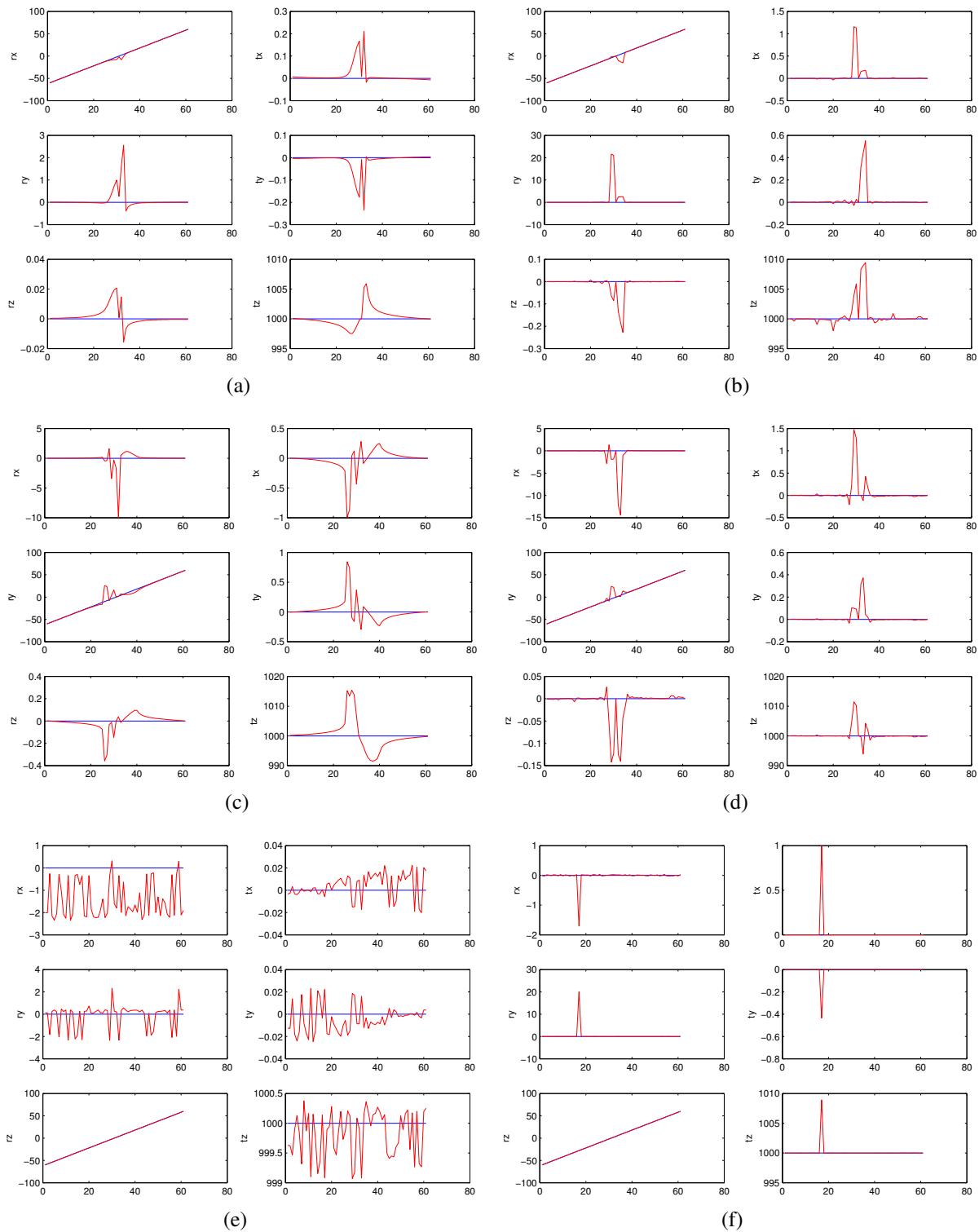


Figura 7.8: Rotación según eje x para Modern POSIT (a) y para Classic POSIT (b), rotación según eje y para Modern POSIT (c) y para Classic POSIT (d) y rotación según eje z para Modern POSIT (e) y para Classic POSIT (f)

CAPÍTULO 8

Filtrado de Kalman para estimación de pose

8.1. Introducción

Luego de obtener la pose de la cámara para un cuadro, se procedió a evaluar el desempeño de la aplicación en tiempo real. Una de las cosas que se notó fue que había un ruido en la estimación. Por ejemplo con el dispositivo quieto se puede ver que la pose varía. Si bien numéricamente no es mucha la variación, perceptivamente se nota. Es por esto que se decidió implementar un filtro de Kalman para suavizar la estimación.

Otra razón por la cual se usó el filtro de Kalman fue para realizar fusión de sensores. El dispositivo cuenta con sensores iniciales y es posible obtener la variación en la orientación entre dos cuadros. Esta información se fusionó con la información obtenida de la pose mediante un filtro de Kalman.

En este capítulo se presenta el filtro de Kalman de forma genérica para luego explicar las implementaciones particulares del caso de Kalman para suavizado y Kalman para fusión de sensores.

8.2. Filtro de Kalman

Sea \mathbf{x}_k un vector de estados de dimensión n que evoluciona de acuerdo a la ecuación

$$\mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{w}_k \quad (8.1)$$

donde \mathbf{F}_k es la matriz de transición de estados de dimensiones $n \times n$ y \mathbf{w}_k es el vector de ruido del proceso, es de tamaño n y se modela como $\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$ en donde \mathbf{Q}_k es la matriz de covarianza de tamaño $n \times n$ [?]. Este ruido sirve para ajustar la confianza que se le tiene al modelo utilizado. Si el modelo utilizado es bueno el ruido de proceso \mathbf{w}_k es chico.

Adicionalmente se tiene el vector de observaciones \mathbf{y}_k formado como

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k \quad (8.2)$$

donde \mathbf{y}_k es de largo m . \mathbf{H}_k es la matriz de medición, relaciona las observaciones con los estados del proceso, es de tamaño $m \times n$. \mathbf{v}_k es el vector de ruido en la medición, es de tamaño m y se modela como $\mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k)$ siendo \mathbf{R}_k la matriz de covarianza de tamaño $m \times m$.

Se tiene entonces un proceso representado por un modelo en variables de estado como

$$\begin{cases} \mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{w}_k \\ \mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k \end{cases}$$

El estado del filtro se representa por dos variables, la estimación del estado *a posteriori* en el instante k , $\hat{\mathbf{x}}_{k|k}$, y la matriz de covarianza del error de estimación *a posteriori*, $\mathbf{P}_{k|k}$.

El filtrado parte de una condición inicial $\hat{\mathbf{x}}_{0|0}$ y $\mathbf{P}_{0|0}$. El proceso de filtrado se realiza iterativamente en dos etapas, una de predicción y una de actualización:

■ Predicción

$$\begin{aligned}\hat{\mathbf{x}}_{k|k-1} &= \mathbf{F}_k \hat{\mathbf{x}}_{k-1|k-1} \\ \mathbf{P}_{k|k-1} &= \mathbf{F}_k \mathbf{P}_{k-1|k-1} \mathbf{F}_k^T + \mathbf{Q}_k\end{aligned}$$

■ Actualización

$$\begin{aligned}\mathbf{K}_k &= \mathbf{P}_{k|k-1} \mathbf{H}_k^T [\mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^T + \mathbf{R}_k]^{-1} \\ \hat{\mathbf{x}}_{k|k} &= \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k [\mathbf{y}_k - \mathbf{H}_k \hat{\mathbf{x}}_{k|k-1}] \\ \mathbf{P}_{k|k} &= (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_{k|k-1}\end{aligned}$$

8.3. Kalman para suavizado

La información que se obtiene de la etapa de estimación de pose son la orientación de la cámara expresada en ángulos de Euler y la traslación. Se busca suavizar el efecto del ruido introducido al estimar la rotación, este ruido es el que mas influye en la calidad de la realidad aumentada. Los estados son por lo tanto los tres ángulos de Euler, ψ , θ y ϕ .

Para implementar el filtro de Kalman es necesario contar con un modelo físico que modele el la evolución de los estados. Como el dispositivo es controlado por una persona, no se puede asumir nada en cuanto al movimiento que va a realizar esta persona. Por lo tanto se toma un modelo de posición constante y se asume que los movimientos que se realizan son pequeños. La matriz de evolución de estados \mathbf{F} se toma como la matriz identidad.

La matriz de covarianza del ruido de proceso \mathbf{Q}_k se toma como una matriz diagonal ya que por como se definió el modelo se deduce que el ruido del proceso para cada estado son independientes.

En este caso las observaciones que se tienen coinciden con los tres estados por lo tanto la matriz de medición \mathbf{H} se toma como la matriz identidad.

La matriz de covarianza del ruido de medición se calculó se obtuvo a través de varias mediciones con imágenes sintéticas. Se le aplico el proceso a varias imágenes y se comparó la orientación obtenida contra la real. A partir de estos datos se computó la matriz \mathbf{R} . Se tiene que

$$\mathbf{R} = \begin{pmatrix} 4,962496 & 4,314506 & -0,045967 \\ 4,314506 & 7,023549 & 0,074892 \\ -0,045967 & -0,074892 & 0,001062 \end{pmatrix}$$

Con estos parámetros se implementó el filtro de Kalman para suavizado.

8.4. Kalman con sensores

Para hacer la fusión de sensores se tomaron los mismos estados, la misma matriz de evolución de estados y el mismo error de proceso. Como observaciones se tienen los ángulos de Euler calculados por POSIT (ψ_C, θ_C, ϕ_C) y los ángulos de Euler a la salida de los sensores inerciales (ψ_S, θ_S, ϕ_S). La ecuación de observaciones de Kalman queda

$$\begin{pmatrix} \psi_C \\ \theta_C \\ \phi_C \\ \psi_S \\ \theta_S \\ \phi_S \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \psi \\ \theta \\ \phi \end{pmatrix} + \begin{pmatrix} 4,962496 & 4,314506 & -0,045967 & 0 & 0 & 0 \\ 4,314506 & 7,023549 & 0,074892 & 0 & 0 & 0 \\ -0,045967 & -0,074892 & 0,001062 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

La matriz de error de medición es una matriz de bloques diagonal compuesta por la matriz de covarianza del error de medición de POSIT y por la matriz de covarianza del error de medición de los sensores. Es razonable tomar la matriz covarianza como diagonal ya que esto implica que el error de medición de POSIT es independiente de el de los sensores. Para los sensores no se contó con ninguna estadística que permitiera calcular la matriz de covarianza, esto es un aspecto a mejorar.

8.5. Resultados

CAPÍTULO 9

Rendering en iOS: ISGL3D

9.1. Introducción

Rendering es un término en inglés que denota el proceso de generar una imagen 2D a partir de un modelo digital 3D o un conjunto de ellos, a los que se les llama “escena”. Puede ser comparado con tomar una foto o filmar una escena en la vida real. Afortunadamente, existen varias herramientas de *rendering*, también llamadas “ motores de juego 3D”, para plataformas móviles, en especial que funcionen sobre iOS. Algunas de ellas son Unity 3D, ISGL3D, Cocos3D, Open GL ES y Shi-Va3D. A continuación serán comentadas tan sólo las consideradas durante el presente proyecto por ser populares y gratuitas.

La primera en ser tomada en cuenta fue “Open Graphics Library Embedded Systems” (Open GL ES), que es un subconjunto de las herramientas de gráficos 3D de Open GL. Fue diseñada para ser utilizada sobre sistemas embebidos (dispositivos móviles, consolas de video juegos, etc.); Open GL es el estándar más ampliamente usado alrededor del mundo para la creación de gráficos 2D y 3D, es gratis y multiplataforma. Como la programación en Open GL y en particular en Open GL ES es de muy bajo nivel y por lo tanto bastante complicada, se optó por investigar otras herramientas. Se descubrió entonces ISGL3D un *framework* escrito en Objective-C que trabaja sobre Open GL ES y que busca facilitar la tarea del programador al momento de crear y manipular escenas 3D mediante una “Application Program Interface” (API) sencilla e intuitiva. Es un proyecto gratis y en código abierto. Luego de algunas semanas de trabajo con la herramienta e importantes avances desde el punto de vista del manejo de la misma se descubrió la existencia de otro *framework* de idénticas características llamado “Cocos3D”. Cocos3D es una extensión de “Cocos2D”, una herramienta para la generación de gráficos 2D, muy popular entre los desarrolladores de aplicaciones para iOS. Como no se identificaron diferencias significativas entre ISGL3D y Cocos3D, se priorizó el tiempo dedicado a ISGL3D y se decidió continuar trabajando de forma inalterada. Al día de hoy, sobre el final del proyecto, se cree que si bien técnicamente ambos *frameworks* son muy buenos y a la vez similares entre sí, ISGL3D parece estar algo más avanzado en cuanto a su desarrollo. Sin embargo debido a la gran popularidad de Cocos2D, Cocos3D ha heredado muchos usuarios y cuenta con una comunidad mucho más activa, lo que facilita mucho su uso y hace pensar que en un futuro cercano resulte en un *framework* más desarrollado.

En este capítulo se comentarán algunas características y conceptos de ISGL3D que fueron importantes para el proyecto; por detalles de algunos temas en particular referirse a la referencia de la ya mencionada API en: www.isgl3d.com/resources/api. Además se trazará una hoja de ruta para todo aquel que quiera iniciarse en el manejo de la herramienta.

9.2. Conceptos básicos de ISGL3D

ISGL3D es un motor de juegos 3D para *iPad*, *iPhone* y *iPod touch* escrito en *Objective-C*, que sirve para crear escenas y *renderizarlas* de forma sencilla. Es un proyecto en código abierto y gratis. En su sitio web oficial: www.isgl3d.com, se puede descargar su código, y de forma sencilla ISGL3D puede ser agregado como un complemento de *Xcode*. Además se pueden encontrar tutoriales, detalles de su API y un acceso a un grupo de *Google* donde la comunidad pregunta y responde dudas propias y ajenas. Una buena manera de iniciarse en manejo de la herramienta es siguiendo los tutoriales en: www.isgl3d.com/resources/tutorials; al menos este fue el camino elegido por el grupo. Los tutoriales son 6, y abarcan distintos tópicos:

- **Tutorial 0:** primer paso en el creado de una aplicación ISGL3D. Cubre algunos conceptos básicos y muestra cómo integrar la herramienta a *Xcode*.
- **Tutorial 1:** muestra cómo crear una escena bien simple, con tan sólo un cubo en rotación continua.
- **Tutorial 2:** enseña cómo agregar luz a una escena. Se ven las distintas fuentes de luz que existen en el *framework*.
- **Tutorial 3:** se ve cómo hacer para mapear texturas en los objetos 3D con el objetivo de hacerlos más realistas.
- **Tutorial 4:** muestra cómo crear interacción entre el usuario y los distintos objetos ISGL3D, cuando este los toca a través de la pantalla.
- **Tutorial 5:** se ven algunas nuevas primitivas (modelos básicos) y se muestra cómo agregar transparencia a los objetos.

Al descargar e instalar ISGL3D, se puede ver que la herramienta incluye un proyecto *Xcode* integrado por varios ejemplos para ejecutar y a la vez ver su código, otra buena forma de aprender cómo realizar distintas tareas de interés. Entre los ejemplos se encuentra la solución a cada uno de los tutoriales.

Cuando se crea una aplicación ISGL3D, el núcleo de la misma es la llamada “view” (“vista” en Español). Una *view* está compuesta principalmente por una escena y una cámara:

- Una **escena** (*Isgl3dScene3D*) es donde los objetos o modelos 3D son agregados como nodos. Todos los nodos pueden ser tanto trasladados como rotados y pueden tener otros nodos hijos; los nodos hijos son trasladados y rotados con sus padres. Así como objetos 3D, se pueden agregar luces de distinto tipo, que generarán en la escena efectos de sombra que luego serán adecuadamente *renderizados* en función de dónde se encuentre y hacia dónde esté mirando la cámara.
- Una **cámara** que es utilizada para ver la escena desde una posición y un ángulo en particular. La cámara se manipula como cualquier otro objeto o nodo en la escena, se puede trasladar, rotar y hasta indicar hacia dónde quiere uno que esta apunte. Es importante ajustar la cámara de manera que su arquitectura sea la que uno busca. Se pueden entonces ajustar ciertos parámetros intrínsecos a esta como por ejemplo su campo visual, su distancia focal, la altura y la anchura del plano imagen, etc.

Es importante entender que el llamado *render* se realiza sumando la información de la escena, objetos 3D y sus hijos, luces, etc.; más la información de dónde se encuentra la cámara, sus características y hacia dónde esta apunta.

9.3. FOV y ejes de ISGL3D

Una particularidad de la cámara de ISGL3D es que el parámetro intrínseco “distancia focal” visto en la sección 6.2, no es directamente configurable. En cambio, el valor que sí se puede alterar es el llamado “FOV”, acrónimo de “Field Of View”. El *field of view* de una cámara no es más que su campo visual, y se mide como la extensión angular máxima mapeable en el plano imagen, medida desde el centro óptico O. Puede ser medido de forma horizontal o de forma vertical; sin embargo, en ISGL3D es definido verticalmente. Ver figura 9.1.

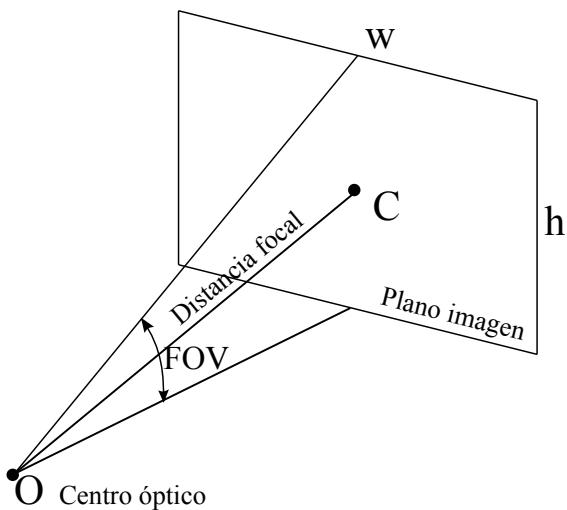


Figura 9.1: Definición gráfica del FOV.

Realizando algo de geometría se ve que la relación entre la distancia focal y el FOV es:

$$FOV = 2 \cdot \text{arctg} \left(\frac{h}{2 \cdot f} \right)$$

donde h denota la altura del plano imagen y f la distancia focal de la cámara.

Otra particularidad de ISGL3D es el sistema de coordenadas que difiere del que se usa en algunas herramientas de modelado, como por ejemplo Blender, en las que el plano X-Y se corresponde con el plano horizontal. Para ISGL3D los ejes están orientados como en la Figura 9.2, con el origen en el centro del plano de la pantalla del dispositivo que coincide con el plano X-Y, y el eje Z saliente de la misma.

9.4. Primitivas de ISGL3D

ISGL3D cuenta con algunas estructuras primitivas que pueden ser usadas como modelos, o incluso combinadas de manera de formar modelos algo más complejos. Las principales estructuras

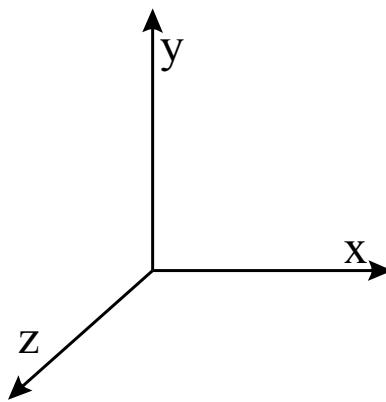


Figura 9.2: Sistema de coordenadas de ISGL3D.

primitivas de ISGL3D son:

- **Isgl3DArrow:** modelo correspondiente a una flecha. Tiene 4 parámetros configurables:

- *headHeight*: altura de la punta.
- *headRadius*: radio de la punta.
- *height*: altura total de la flecha.
- *radius*: radio de la base.

- **Isgl3DCone:** modelo correspondiente a un cono. Tiene 3 parámetros configurables:

- *bottomRadius*: radio de la base inferior.
- *height*: altura del cono.
- *topRadius*: radio de la base superior.

- **Isgl3DCube:** modelo correspondiente a un cubo. Tiene 3 parámetros configurables:

- *depth*: profundidad del cubo.
- *height*: altura del cubo.
- *width*: anchura del cubo.

- **Isgl3DCylinder:** modelo correspondiente a un cilindro. Tiene 3 parámetros configurables:

- *height*: altura del cilindro.
- *radius*: radio del cilindro.
- *openEnded*: indica si el cilindro cuenta con sus extremos abiertos o no.

- **Isgl3DEllipsoid:** modelo correspondiente a una elipsoide. Cuenta con 3 parámetros configurables:

- *radiusX*: radio de la elipsoide en la dirección *x*.
- *radiusY*: radio de la elipsoide en la dirección *y*.
- *radiusZ*: radio de la elipsoide en la dirección *z*.

- **Isgl3DOvoid:** modelo ovoide. Cuenta con 3 parámetros configurables:
 - a : radio del ovoide en la dirección x .
 - b : radio del ovoide en la dirección y .
 - k : factor que modifica la forma de la curva. Cuando toma el valor 0, el modelo se corresponde con el de una ellipsoide.
- **Isgl3DSphere:** modelo correspondiente a una esfera. Tiene un único parámetro configurable:
 - $radius$: radio de la esfera.
- **Isgl3DTorus:** modelo correspondiente a un toroide. Cuenta con 2 parámetros configurables:
 - $radius$: radio desde el origen del toroide hasta el centro del tubo.
 - $tubeRadius$: radio del tubo del toroide.

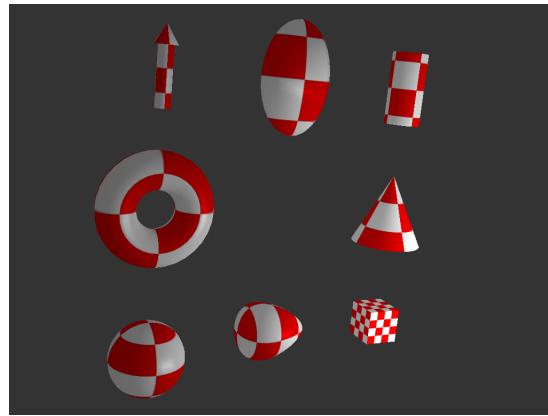


Figura 9.3: Principales primitivas en ISGL3D.

Para la creación de cada primitiva, se debe especificar además, la cantidad de segmentos que la forman en las distintas dimensiones. En la figura 9.3 se pueden ver todas las primitivas anteriores. Es fácil ver que dichas primitivas cuentan con cierta textura cuadriculada de colores rojo y blanco, que fue lograda mapeando una imagen sobre cada una de ellas. La porción de código que se usó para realizar tal mapeo se muestra a continuación:

```
Isgl3dTextureMaterial * material = [Isgl3dTextureMaterial
                                     materialWithTextureFile:@"red_checker.png" shininess:0.9];

Isgl3dTorus * torusMesh = [Isgl3dTorus meshWithGeometry:2 tubeRadius:1 ns:32 nt:32];

Isgl3dMeshNode * _torus = [self.scene createNodeWithMesh:torusMesh andMaterial:material];
```

En la primera línea de código se crea el material. Dicho material es del tipo *Isgl3dTextureMaterial*; y la imagen con la que este se crea es la de la figura 9.4. Luego, se crea el toroide asignándole los parámetros vistos más atrás en esta sección; y finalmente, se crea y se agrega a la escena el nodo asociado al toroide, con el material creado al principio.

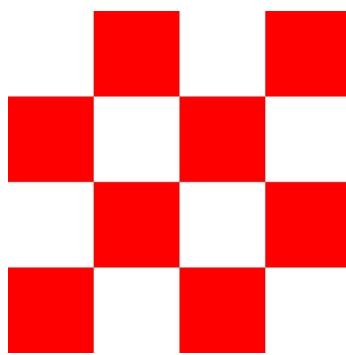


Figura 9.4: Imagen *red_checker.png*, utilizada para crear la textura asociada a las primitivas de la figura 9.3.

9.5. Importación de modelos a ISGL3D.

A veces lo que se quiere no es agregar a la escena una primitiva sino un modelo previamente creado. Los modelos son realizados en herramientas de creado y animación de gráficos 3D como por ejemplo *Blender*, *MeshLab*, *Autodesk Maya* o *Autodesk 3ds Max*. Luego deben ser exportados en un formato llamado *COLLADA*, acrónimo de “COLLABorative Design Activity”, que sirve para el intercambio de contenido digital 3D entre distintas aplicaciones de modelado. Por su parte, ISGL3D permite importar modelos pero en un formato llamado “POD”. Se usó entonces, una aplicación llamada *Collada2POD* que lo que hace es convertir modelos tridimensionales en formato *COLLADA* al formato POD. *Collada2POD* puede ser descargada gratuitamente de la página oficial de *Imagination Technologies*, su desarrollador: <http://www.imgtec.com>.

Una vez que se tiene al objeto 3D en el formato requerido, este puede ser importado en ISGL3D de forma sencilla:

```
Isgl3dPODImporter * podImporter = [Isgl3dPODImporter podImporterWithFile:@“modelo.pod”];  
Isgl3dNode * _model = [self.scene createNode];  
[podImporter addMeshesToScene:_model];  
_model.position = iv3(2, 6, 0);
```

En la primera línea de código se instancia la clase *Isgl3dPODImporter* que sirve para transformar modelos POD a objetos ISGL3D, y se le asigna a la misma el modelo “modelo.pod”. Luego, se crea un nodo llamado “_model”, al que se le asigna el modelo; y se agrega a la escena. Finalmente, se le asigna al nodo una posición. En la figura 9.5 se puede ver un modelo de José Artigas, agregado dos veces a una misma escena, pero visto desde ángulos distintos.

Si lo que se quiere es que los modelos sean animados, o lo que es lo mismo, que tengan movimiento, hay dos soluciones posibles a tomar en consideración:

- **Modelo animado:** muchas veces lo que se tiene es un modelo 3D animado desde su construcción. Estos pueden ser creados, al igual que los modelos 3D inanimados, con las herramientas para el creado y la animación de gráficos 3D antedichas. Existe mucha bibliografía al respecto, además de haber múltiples sitios en internet de donde bajar los modelos, incluso en forma gratuita. Luego de obtenido el modelo animado, lo que se tiene es precisamente al modelo, pero con una línea de tiempo con las animaciones. Nuevamente, hay que convertirlo a formato POD para ser usado en ISGL3D. El código para poder visualizar al modelo es:

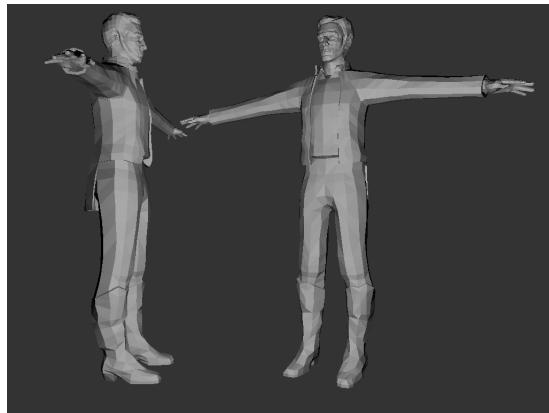


Figura 9.5: Modelo de José Artigas agregado dos veces a una misma escena, pero visto desde ángulos distintos.

```
Isgl3dPODImporter * podImporter = [Isgl3dPODImporter
    podImporterWithFile:@"animated_model.pod"];

Isgl3dSkeletonNode *_model = [self.scene createSkeletonNode];

[podImporter addMeshesToScene:_model];

Isgl3dAnimationController *_animationController = [[Isgl3dAnimationController alloc]
    initWithSkeleton:_model andNumberOfFrames:[podImporter numberOfFrames]];

[_animationController start];
```

En la primera línea de código se instancia la clase *Isgl3dPODImporter*, y se le asigna a la misma el modelo animado “animated_model.pod”. Luego, se crea y se agrega a la escena un nodo del tipo *Isgl3dSkeletonNode* llamado “_model” que contiene al modelo animado. La clase *Isgl3dSkeletonNode* provee una interfaz sencilla para animar al ahora objeto ISGL3D, que con la ayuda de la clase *Isgl3dAnimationController*, logra automatizar el movimiento del mismo. Finalmente, se instancia y configura la clase *Isgl3dAnimationController* y se le da inicio a la animación en la última línea.

- **Múltiples modelos inanimados:** otra forma de animar un modelo 3D es usando múltiples modelos inanimados. Estos pueden ser cargados en ISGL3D como un único objeto o nodo y mediante ciertas instrucciones sencillas, se le dice al *framework* que presente uno a continuación del otro, interpolando entre posiciones contiguas, lo que genera una sensación de movimiento. Este fue el método utilizado en el presente proyecto para animar al modelo de José Artigas. En la figura 9.6 se puede ver al mismo en 3 posiciones distintas.

El código para realizar la interpolación mencionada, aplicado por ejemplo a dos modelos, será:

```
Isgl3dPODImporter * podImporter = [Isgl3dPODImporter
    podImporterWithFile:@"model_1.pod"];

[podImporter buildSceneObjects];

Isgl3dPODImporter * podImporter2 = [Isgl3dPODImporter
```



Figura 9.6: Modelo de José Artigas en 3 posiciones distintas, utilizadas para generar en ISGL3D una sensación de movimiento.

```

podImporterWithFile:@"model_2.pod"];

[podImporter2 buildSceneObjects];

Isgl3dGLMesh* _modelMesh = [podImporter meshAtIndex:0 ];

Isgl3dGLMesh* _modelMesh2 = [podImporter2 meshAtIndex:0 ];

Isgl3dKeyframeMesh * _mesh = [Isgl3dKeyframeMesh keyframeMeshWithMesh:_modelMesh];

[_mesh addKeyframeMesh:_modelMesh2];

[_mesh addKeyframeAnimationData:0 duration:1.0f];
[_mesh addKeyframeAnimationData:0 duration:2.0f];
[_mesh addKeyframeAnimationData:1 duration:1.0f];
[_mesh addKeyframeAnimationData:1 duration:2.0f];

[_mesh startAnimation];

Isgl3dNode * node = [_container createNodeWithMesh:_mesh
                    andMaterial:[podImporter materialWithName:@"material_0"]];

node.position = iv3(-90, -60, -150);

[podImporter addMeshesToScene:node];

```

En las primeras 4 líneas de código se instancia en dos oportunidades la clase *Isgl3dPODImporter*, y se les asigna a las instancias los modelos inanimados “model_1.pod” y “model_2.pod”. La instrucción *buildSceneObjects* crea todos los objetos de la escena del modelo POD, pero no los agrega a la escena ISGL3D. Luego se obtienen los modelos indexados de cada uno de los PODs (cada POD puede tener una escena con más de un modelo, mediante un índice se referencia qué modelo se quiere obtener) y se almacenan en “_modelMesh” y “_modelMesh2” respectivamente. Se genera a continuación un nuevo modelo al que se le asignan los dos modelos anteriores, luego se programa la animación y se le da inicio mediante la instrucción *startAnimation*. Finalmente, se genera un nuevo objeto o nodo ISGL3D al que se le asigna el modelo, y un material también cargado desde el archivo POD; se le asigna además una posición y se lo agrega a la escena.

9.6. Luz en ISGL3D

Un tema importante al momento de proyectar modelos 3D en una escena es la luz. La visualización de un modelo puede cambiar significativamente en función de las características de luz que tenga una escena. En ISGL3D la se logra mediante la suma de tres componentes independientes: *ambiente*, *difusa* y *especular*. La luz ambiente es no direccional y está presente en toda la escena. La luz difusa ya implica la reacción que tiene la luz proveniente de las distintas fuentes de luz direccionales sobre las superficies de los objetos que existen en la escena generando haces de luz en direcciones aleatorias. La luz specular modela el comportamiento de la luz reflejada sobre las distintas superficies en ciertas direcciones particulares (no aleatorias) que dependen del coeficiente de reflexión de los materiales. Cada uno de estos tres tipos de luz que modelan el mundo real, existen en ISGL3D y son representados como características configurables de los objetos de luz. A su vez existen fuentes lumínicas de distintos tipos: *puntual*, *direccional* y *cónica*. Cada tipo tiene una función distinta de la atenuación con respecto a la distancia. Un ejemplo de la creación de un elemento lumínico para una escena se ve en el siguiente código:

```
Isgl3dLight * _redLight = [Isgl3dLight lightWithHexColor:@“FF0000” diffuseColor:@“FF0000” specularColor:@“FFFFFF” attenuation:0.02];
_redLight.renderLight = YES;
[self.scene addChild:_redLight];
```

En la primera línea de código se crea una luz con una componente ambiente de color rojo, una componente difusa también de color rojo y una componente specular de color blanca. En todos los casos se usa la notación hexadecimal de las componentes RGBA. Además se escogió un valor 0,02 de atenuación. Luego, se pide que el foco de luz sí sea *renderizado* y finalmente, este es agregado a la escena. Por defecto, el tipo de luz utilizado es puntual.

9.7. Método - (*void*) *tick:(float)dt*

Cuando se instancia la clase encargada de generar el render (clase *HelloWorldView*), la misma ejecuta la configuración básica de inicialización del objeto. Entre otras cosas, dentro del código de inicialización, se agrega lo siguiente:

```
[self schedule:@selector(tick:)];
```

Esto lo que hace es “agendarse” una invocación del método *tick* en forma periódica. Dentro de dicho método es que se hace la actualización de la escena, y el en caso particular de este proyecto, es en dónde se actualiza la posición de los objetos 3D en función de la posición de la cámara respecto de los ejes del mundo. Ver capítulo 6. Este método es de vital importancia por tratarse de uno de los dos *callbacks* que toda aplicación de realidad aumentada tiene (el otro es la captura y procesamiento de la imagen para obtener la pose).

9.8. ISGL3D en la aplicación

En el capítulo 11 se explican algunos detalles sobre el uso de ISGL3D dentro de la aplicación final. En particular se dan detalles constructivos sobre cómo utilizar esta herramienta para generar *renders* sobre un fondo que sea la captura de la cámara y de la convivencia de los dos *callbacks* de la aplicación.

9.9. Conclusión

En el presente capítulo se mencionaron algunas herramientas existentes para realizar *renders* en dispositivos móviles con sistema operativo iOS. Luego, se justificó la elección de ISGL3D para tal función y se dieron algunos conceptos introductorios a la herramienta; además se dió una hoja de ruta para todo aquel que le interese ampliar sus conocimientos en el *framework*. Se indicaron dos formas distintas de embeber un modelo 3D en ISGL3D y se aclaró cuál fue la forma escogida por el grupo. Finalmente, se referenció el capítulo 11, en donde se explica cómo se usó ISGL3D en la aplicación final.

CAPÍTULO 10

Casos de Uso

10.1. Introducción

En este capítulo se presentan los distintos casos de uso que se implementaron con el fin de integrar los algoritmos comentados en capítulos anteriores en pequeñas aplicaciones que funcionen *de punta a punta*. Se buscó resolver individualmente los diferentes desafíos técnicos que una aplicación real de realidad aumentada para museos puede llegar a tener. Estas últimas no serán más que una combinación guionada de cada uno de estos casos de uso.

A lo cargo del capítulo se verán entonces los tres casos de uso implementados: “interactividad”, “video” y “modelos”. El primero presenta un modelo simple sobre el marcador que responde a toques con cierto movimiento y un audio en particular, el segundo soluciona el problema de proyectar un video sobre el marcador de forma consistente con el movimiento del usuario. El último caso de uso muestra cómo es posible importar modelos a ISGL3D de manera de lograr realidades aumentadas mucho más interesantes que si tan sólo se hicieran con las primitivas del *framework*, por detalles ver capítulo 9.

10.2. Caso de uso “interactividad”

10.2.1. Comentarios sobre el caso de uso

En este caso de uso se implementa la parte interactiva de la aplicación. Al enfocar el marcador, se puede ver un cubo sobre el QISet de la esquina superior izquierda. Ver Figura 10.1. Si el cubo es tocado a través de la pantalla del dispositivo, este se anima y se reproduce un audio que indica la posición del cubo en el instante de ser presionado. Inmediatamente después, es desplazado hacia el QISet de la esquina superior derecha. Nuevamente, si el cubo es tocado a través de la pantalla del dispositivo, este se anima y se reproduce un audio que indica la posición del cubo en el instante de ser presionado. Inmediatamente después, este se desplaza hacia el QISet restante. Lo anterior sucederá de forma cíclica, cada vez que se presione sobre el mpdelo.

Esta funcionalidad es fundamental si lo que se quiere implementar es por ejemplo una audioguía interactiva. Podría pensarse una aplicación en la que el cubo anterior se reemplace por flechas 3D, y que estas sean ubicadas conjuntamente en distintas partes de una obra. Entones, al seleccionar cada una de las flechas, se podría reproducir un audio con información referente a esa zona o punto en particular.

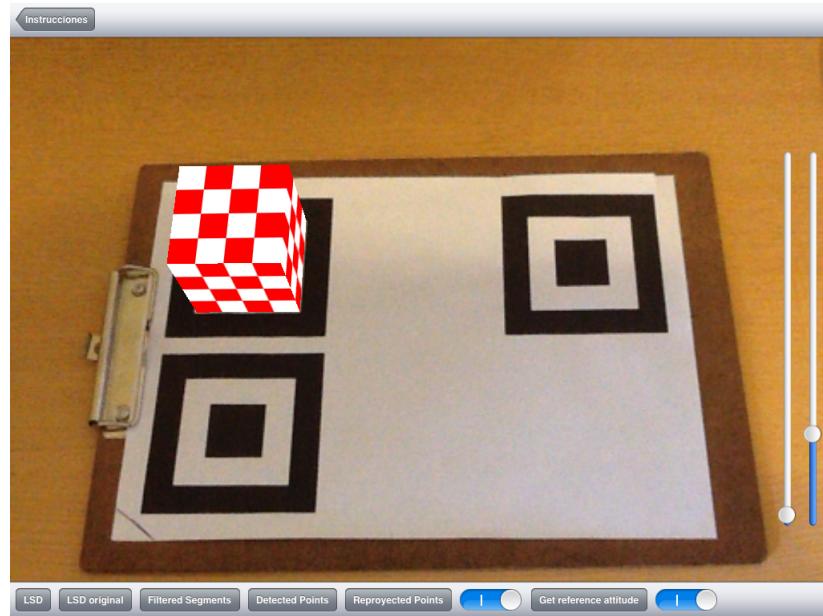


Figura 10.1: Captura de pantalla del caso de uso “interactividad”. Se puede ver al cubo apoyado sobre el QLSet de la esquina superior izquierda y los diferentes controles que ayudan a la depuración del código.

Esta aplicación también se utilizó con fines de *debugging* o depuración de la integración de cada uno de los bloques. Se le agregaron las siguientes funcionalidades:

- La posibilidad de ver dibujados sobre la imagen los segmentos detectados por LSD. En sus versiones original y optimizada.
- La posibilidad de ver dibujados sobre la imagen los segmentos filtrados pertenecientes al marcador. Así como también las esquinas detectadas de cada uno de los cuadriláteros que lo forman.
- La posibilidad de variar el umbral utilizado para el filtrado de segmentos.
- La posibilidad de ver las esquinas de cada uno de los cuadriláteros que forman al marcador reproyectadas según la pose del dispositivo obtenida.
- La posibilidad de prender o apagar el filtro de Kalman.
- La posibilidad de aumentar o disminuir el ruido de medición del filtro de Kalman.
- La posibilidad de elegir si usar o no la fusión de la estimación de pose con los sensores.

En la Figura 10.1 también se puede ver cómo es la interfaz de usuario de este caso de uso, en donde se puede elegir entre todas las funcionalidades anteriores. El mismo fue fundamental para evaluar el desempeño de los algoritmos utilizados funcionando en tiempo real. Gracias a estas funcionalidades se pudieron definir las condiciones para las cuales el conjunto de todos los bloques funciona mejor. Se fue variando la distancia al marcador y se ajustó el umbral para el filtro de segmentos. Además, se pudieron ajustar los parámetros del filtro de Kalman y se pudo comparar el desempeño de la estimación de pose utilizando solamente información de la cámara con el resultado obtenido de la fusión de sensores. Fue en este caso de uso que se evaluó cualitativamente el

desempeño de la versión optimizada de LSD, respecto del de la versión original.

Si bien todas estas pruebas y ajustes si hicieron previamente en una computadora y con imágenes de prueba, fue necesario contar con una aplicación en la que se pudiera ver sobre el dispositivo al conjunto de los algoritmos funcionando en tiempo real.

10.2.2. Detalles constructivos

10.2.2.1. Objetos ISGL3D interactivos

La manera de agregar interactividad a un nodo ISGL3D es bastante sencilla. En primer lugar, debe configurarse su propiedad *interactive* de forma positiva y luego se le debe ejecutar el método *addEvent3DListener*:

```
Isgl3dTextureMaterial * material = [Isgl3dTextureMaterial
materialWithTextureFile:@"red_checker.png" shininess:0.9
precision:Isgl3dTexturePrecisionMedium repeatX:NO repeatY:NO];

Isgl3dCube* cubeMesh = [Isgl3dCube meshWithGeometry:60 height:60 depth:60 nx:40 ny:40];

Isgl3dNode * _cubito = [self.scene createNodeWithMesh:cubeMesh andMaterial:material];

_cubito.interactive =YES;

[_cubito addEvent3DListener:self method:@selector(objectTouched:) forEventType:TOUCH_EVENT];
```

En el código anterior, primero se crea un nodo llamado “_cubito” con la primitiva de un cubo y cierto material. Luego, se indica que sí se quiere que dicho nodo tenga interactividad y finalmente se lo configura para que cuando “_cubito” reciba eventos del tipo *TOUCH_EVENT*, o lo que es lo mismo, cuando se lo toque; se ejecute el método *objectTouched*, definido en la misma clase que esta escrita el código (*self*).

En este caso de uso lo que se hizo en *objectTouched* no fue más que cambiar la posición del cubo en la escena y reproducir un audio dependiente de la posición del mismo.

10.2.2.2. Reproducción de audio en Objective-C

Para reproducir audios en Objective-C primero la clase en la que se quiere reproducir el audio debe importar el framework *AVFoundation* y luego debe implementar el protocolo *AVAudioPlayer-Protocol*. El código que se debe escribir es el siguiente:

```
NSURL *url =[NSURL fileURLWithPath:[NSString stringWithFormat:@"%@/ %@",
[[NSBundle mainBundle] resourcePath],audio.mp3]];

AVAudioPlayer * audioPlayer =[[AVAudioPlayer alloc] initWithContentsOfURL:url error:nil];

audioPlayer.numberOfLoops=0;

audioPlayer.delegate = self;

[audioPlayer play];
```

En la primera línea se genera un *url* que indica cuál es el audio a reproducir y luego se le asigna a una instancia de la clase *AVAudioPlayer*. Se dice que no se quiere reproducir el audio en bucle,

se asigna a la clase en la que se esta escribiendo el código como la delegada de *audioPlayer*, una instancia de *AVAudioPlayer*, y finalmente se le da inicio al audio. Luego de reproducido el audio, se ejecuta autamáticamente el método de firma:

```
- (void)audioPlayerDidFinishPlaying:(AVAudioPlayer *)player successfully:(BOOL)flag;
```

En este código es en donde se indica que la próxima vez que se presione sobre el cubo, se querrá reproducir un audio distinto.

10.2.2.3. Dibujar en ISGL3D

Lo que se hizo fue crear una clase nueva, del tipo *UIView*, a la que se la llamó “claseDibujar”. Esta fue agregada como *subView* de la *view* en donde se muestra el video por detrás de lo que dibuja ISGL3D. Dicha clase se configuró para que fuera transparente y del mismo tamaño que la pantalla del *iPad*. *claseDibujar* cuenta con una cantidad de propiedades a las que se les asignan los diferentes puntos o segmentos que se quieren dibujar; son del tipo “puntero a entero” y “puntero a *float*” respectivamente. Luego, un método llamado *drawRect* es el que se encarga de dibujar cada uno de los puntos y segmentos. Los puntos se dibujan con las siguientes líneas de código:

```
CGContextRef context = UIGraphicsGetCurrentContext();  
  
CGContextStrokeRect(context, CGRectMake(punto_X, punto_Y, 4, 4));
```

En la primera línea de código se crea un contexto. Un contexto contiene ciertos parámetros y toda la información específica del dispositivo, requerida para poder dibujar. En la segunda línea se dibuja cada punto como un rectángulo centrado en el punto en cuestión y con 4 píxeles de ancho y largo. Los segmentos se dibujan con las siguientes líneas de código:

```
CGContextRef context = UIGraphicsGetCurrentContext();  
  
CGContextStrokeLineSegments(context, puntos, 2);
```

En la primera línea de código se crea un contexto (este paso puede saltarse si ya fue creado anteriormente), y en la segunda se dibuja la línea. La variable “puntos” es un arreglo de dos variables del tipo *CGPoint*, cada una de ellas tiene dos valores en precisión simple correspondientes a las coordenadas de un punto. Además, se le configura al segmento una anchura de 2 píxeles.

Finalmente, es bueno aclarar que *claseDibujar* se instancia y se destruye cuadro a cuadro; el método *drawRect* se invoca cada vez que se instancia la clase.

10.3. Caso de uso “video”

10.3.1. Comentarios sobre el caso de uso

Este caso de uso proyecta un video sobre el *QlSet* de la esquina superior izquierda del marcador de manera consistente con la pose del dispositivo. Ver Figura 10.2. La aplicación de esta solución técnica es directa. Tan sólo ajustando un par de parámetros el video podría ser proyectado dentro del marco de un cuadro, sobre uno de sus extremos, sobre una pared blanca o incluso sobre un mapa. Esto puede ser de gran interés para un museo, por ejemplo como complemento a una audioguía. A continuación se explican brevemente algunos detalles técnicos que fue necesario solucionar para lograr implementar este caso de uso.



Figura 10.2: Captura de pantalla del caso de uso “video”. Se puede ver al video proyectado sobre el QlSet de la esquina superior izquierda

10.3.2. Detalles constructivos

Para lograr lo propuesto para este caso de uso se implementó un proyecto que proyecta el video en uno de los cuadrados del marcador. De esta manera, de toda la lógica de estimación de pose, solamente se hace uso de la detección y filtrado. En particular no se hace uso de los resultados del posit. Teniendo entonces detectados los cuatro puntos en los que se quiere reproducir el video parecería que el problema está resuelto. Sin embargo, xcode no permite posicionar en forma directa una vista de video en cualquier conjunto de cuatro puntos.

Si simplemente se quiere reproducir un video, y no se quiere procesar el contenido, lo más cómodo para hacerlo es utilizar la clase *MPMoviePlayerController* que hereda de *NSObject*. Una alternativa similar es haciendo uso de la clase *MPMoviePlayerViewController* que hereda de *UIViewController* y tiene como única propiedad una del tipo *MPMoviePlayerController*.

MPMoviePlayerController tiene un atributo *view* del tipo *UIView* que es la vista y es este atributo el que se quiere posicionar en los cuatro puntos detectados por el filtro. Un atributo del tipo *UIView* tiene un atributo *frame* que es del tipo *CGRect*

```
theMovie.view.frame = CGRectMake(0, 0, 60, 60);
```

En el código anterior *theMovie* es del tipo *MPMoviePlayerController*. De esta manera, se tiene el inconveniente de que en principio cualquier video parecería que solamente puede ser reproducido sobre rectángulos y no en cualquier polígono de cuatro puntos por ejemplo. Sin embargo algo que sí se puede hacer a las instancias de la clase *UIView* es una transformación afín o incluso, de manera más genérica, una homografía.

10.3.3. *CGAffineTransform* y *CATransform3D*

La clase *UIView* tiene una propiedad llamada *transform* que es del tipo *CGAffineTransform*. Las primeras letras de esta clase (CG) refieren a la API **Core Graphics** utilizada ampliamente como herramienta para resolver *rendering* y cualquier tipo de transformación en 2D.

La clase *UIView* también tiene una propiedad llamada *layer* que es del tipo *CALayer* y que permite

realizar transformaciones del tipo *CATransform3D*. Las primeras letras de estas dos clases (*CA*) refieren a la API **Core Animation** que es utilizada para generar animaciones y transformaciones sobre objetos 3D solamente indicando un punto inicial y final para el objeto (también es posible agregar efectos para la transición). En definitiva para resolver el problema del caso de uso existen a priori dos alternativas posibles: *CGAffineTransform* y *CATransform3D*.

Se pueden generar fácilmente instancias transformaciones afines invocando la siguiente función:

```
CGAffineTransform CGAffineTransformMake (
    CGFloat a,
    CGFloat b,
    CGFloat c,
    CGFloat d,
    CGFloat tx,
    CGFloat ty
);
```

que toma 6 *CGFloats* y crea una *CGAffineTransform*, donde cada uno de los valores anteriores se corresponde con los elementos de una matriz transformación afín de la siguiente manera:

$$\begin{pmatrix} a & b & 0 \\ c & d & 0 \\ tx & ty & 1 \end{pmatrix}$$

Así entonces, de los 9 valores de la matriz, 2 de ellos son nulos por tratarse de una transformación afín y otro de ellos es unitario como valor de escala. Resolviendo el sistema como se muestra en la sección 10.3.4 y obteniendo los restantes 6 valores, se le puede asignar transformaciones a la propiedad *transform* y realizar la trasnformación deseada. Este método tuvo como inconveniente el hecho de que

EXPLICAR POR QUE NO FUNCIONÓ

Por su parte también es sencillo generar instancias de transformaciones 3D debido a que existe el tipo de dato definido para generar la matriz *CATransform3D* como:

```
struct CATransform3D
{
    CGFloat m11, m12, m13, m14;
    CGFloat m21, m22, m23, m24;
    CGFloat m31, m32, m33, m34;
    CGFloat m41, m42, m43, m44;
};

typedef struct CATransform3D CATransform3D;
```

donde m_{ij} corresponde al elemento de la matriz ubicado en la fila i columna j . Así entonces también es posible, conociendo los valores de la homografía, completar los elementos de esta matriz 4x4 y asignársela a la propiedad *layer* de la *UIView*. Esta opción de generar una transformación 3D permite incluir transformaciones más generales que una homografía o una transformación afín. Si lo que se busca es que esta matriz represente una homografía (2D), es necesario entonces que la coordenada z sea nula, es decir

$$\begin{pmatrix} m_{11} & m_{12} & 0 & m_{14} \\ m_{21} & m_{22} & 0 & m_{24} \\ 0 & 0 & 1 & 0 \\ m_{41} & m_{42} & 0 & m_{44} \end{pmatrix}$$

donde a su vez m_{44} se asume de valor unitario por ser un factor de escala. De la misma manera que para la transformación afín, resolviendo la homografía como se ve en la sección 10.3.4 se obtienen los 8 valores restantes de la matriz y es posible asignarle una homografía a un objeto *UIView* para resolver el problema presente.

10.3.4. Resolución de Homografía

A continuación se hace el desarrollo de la resolución del sistema de ecuaciones que se tuvo que resolver para hallar los parámetros de la homografía que transforma una imagen de referencia en la imagen que se tiene en cada momento como resultado de la captura de la cámara. Se asume entonces que se conocen los puntos de referencia y los puntos de referencia transformados (los detectados luego del filtrado de segmentos) y lo que se quiere averiguar es la matriz h que logra dicha transformación. Esta homografía 2D-2D se puede expresar en forma matricial, en coordenadas homogéneas de la siguiente manera:

$$\begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} i \\ j \\ k \end{pmatrix}$$

donde la matriz h_{3x3} representa la transformación homográfica, el vector $(x, y, z)^t$ representa los puntos de referencia a ser transformados y el vector $(i, j, k)^t$ respresenta los puntos detectados cuadro a cuadro como las esquinas del marcador.

Asumiendo un valor unitario para las coordenadas z y k la resolución del sistema se simplifica mucho y no se pierde generalidad. Imponiendo esto entonces, el sistema anterior se puede expresar de la siguiente forma:

$$xh_{11} + yh_{12} + h_{13} = i \quad (10.1)$$

$$xh_{21} + yh_{22} + h_{23} = j \quad (10.2)$$

$$xh_{31} + yh_{32} + h_{33} = 1 \quad (10.3)$$

Multiplicando la ecuación (10.3) por i e igualándola a la ecuación (10.1) se obtiene lo siguiente:

$$xh_{11} + yh_{12} + h_{13} = ixh_{31} + iyh_{32} + ih_{33} \quad (10.4)$$

o lo que es lo mismo:

$$xh_{11} + yh_{12} + h_{13} - ixh_{31} - iyh_{32} - ih_{33} = 0 \quad (10.5)$$

Procediendo de manera análoga y multiplicando la ecuación (10.3) por j e igualándola a la ecuación (10.2) se obtiene lo siguiente:

$$xh_{21} + yh_{22} + h_{23} = jxh_{31} + jyh_{32} + jh_{33} \quad (10.6)$$

o lo que es lo mismo:

$$xh_{21} + yh_{22} + h_{23} - jxh_{31} - jyh_{32} - jh_{33} = 0 \quad (10.7)$$

Las ecuaciones (10.3) y (10.7) se pueden expresar en forma matricial, de la siguiente manera:

$$\begin{pmatrix} x & y & 1 & 0 & 0 & 0 & -ix & -iy & -i \\ 0 & 0 & 0 & x & y & 1 & -jx & -jy & -j \end{pmatrix} \begin{pmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \\ h_{33} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Teniendo entonces 4 parejas de puntos referencia y puntos transformados y asumiendo h_{33} de valor unitario se tiene entonces 8 ecuaciones y 8 incógnitas, lo que lo vuelve un sistema compatible determinado que se puede expresar de la siguiente manera:

$$\begin{pmatrix} x_0 & y_0 & 1 & 0 & 0 & 0 & -i_0x_0 & -i_0y_0 \\ 0 & 0 & 0 & x_0 & y_0 & 1 & -j_0x_0 & -j_0y_0 \\ x_1 & y_1 & 1 & 0 & 0 & 0 & -i_1x_1 & -i_1y_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -j_1x_1 & -j_1y_1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -i_2x_2 & -i_2y_2 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -j_2x_2 & -j_2y_2 \\ x_3 & y_3 & 1 & 0 & 0 & 0 & -i_3x_3 & -i_3y_3 \\ 0 & 0 & 0 & x_3 & y_3 & 1 & -j_3x_3 & -j_3y_3 \end{pmatrix} \begin{pmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \end{pmatrix} = \begin{pmatrix} i_0 \\ j_0 \\ i_1 \\ j_1 \\ i_2 \\ j_2 \\ i_3 \\ j_3 \end{pmatrix}$$

Así entonces, lo que se hace para resolver la homografía es cuadro a cuadro tener detectados los puntos en los que se quiere presentar la vista del video que se corresponden con cuatro puntos detectados por el filtro y tener las correspondencias con el marcador real, se posiciona la vista en la posición de referencia y se le aplica la homografía hallada que vincula la posición referencia con los puntos detectados.

10.4. Caso de uso “modelos”

10.4.1. Comentarios sobre el caso de uso

10.4.2. Detalles constructivos

CAPÍTULO 11

Implementación

11.1. Introducción

En este capítulo se muestra la integración de los conocimientos adquiridos a lo largo del proyecto para poder llevar a cabo la realidad aumentada en una aplicación real. Si bien el objetivo principal del proyecto era la exploración de distintos métodos y algoritmos, parecía importante poder poner en práctica todo lo desarrollado en un producto final que pudiera parecerse a un prototipo de aplicación comercial. En particular se desarrolló una aplicación pensando en los cuadros de la planta baja del Museo Nacional de Artes Visuales (MNAV). Entre otros autores, tiene cuadros de Pedro Figari, Juan Manuel Blanes y de Joaquín Torres García, que se eligieron para hacer el prototipo. La aplicación consta de distintas funcionalidades tales como:

- (1) Detección QR
- (2) Navegación por listas de cuadros
- (3) Comunicación con un servidor con la base de datos.
- (4) Detección SIFT para identificar el cuadro.
- (5) Diferentes realidades aumentadas según la obra.

En las próximas secciones se describe más en detalle cada uno de estos puntos y su integración a la aplicación final. También se describe el flujo de la aplicación y algunas clases implementadas.

11.2. Diagrama global de la aplicación

En la descripción de las clases que conforman los bloques principales de la aplicación se hace referencia a conceptos de desarrollo sobre Objective-C, así como también a *frameworks* y herramientas utilizadas que fueron explicadas en el capítulo 1. Para la comprensión del detalle de la implementación es importante conocer estos conceptos de desarrollo.

Para que sea más sencilla la comprensión de los bloques que componen la aplicación, en la Figura 11.1 se muestra un diagrama esquemático de la misma que sirve para visualizar cómo es su flujo *a nivel de usuario*.

Al comenzar el recorrido, el usuario tiene la opción de elegir cómo recorrer el museo: de manera *autónoma* o de manera *automática*. En la opción autónoma el usuario es el encargado de elegir

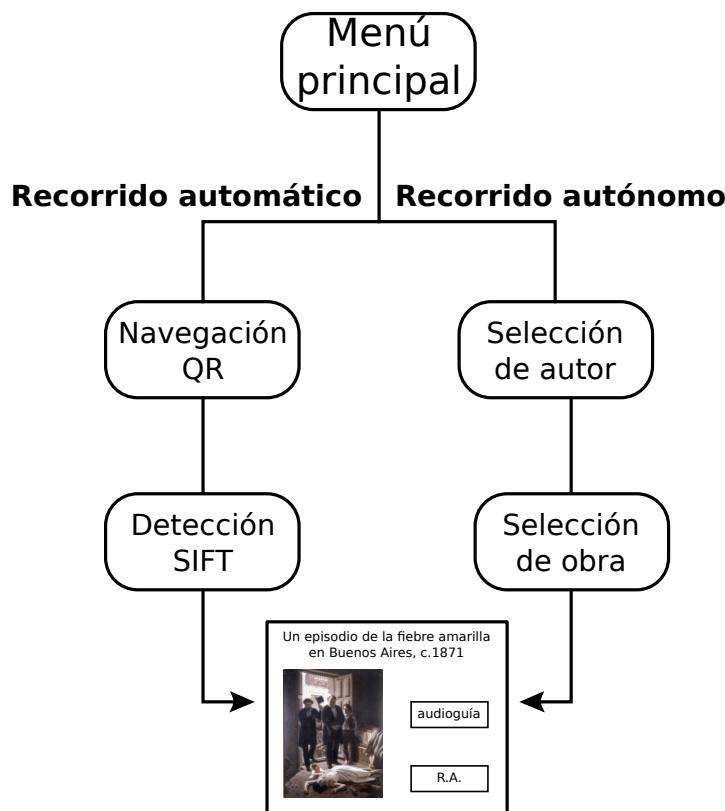


Figura 11.1: Diagrama global de la aplicación

dentro de una lista de autores el que más le interese, y dentro de la lista de cuadros del autor seleccionado, la obra que desea contemplar en detalle. De esta manera el usuario llega eligiendo opciones al cuadro de interés y está listo para comenzar la interacción con la obra, a través de audioguías o realidad aumentada. De la otra manera de recorrer el museo, con la opción automática, el usuario tiene la opción de leer códigos QR desplegados en las distintas salas o secciones del museo, que sirven para identificar en qué parte del museo se encuentra el usuario. De esta manera una vez que el usuario lee el QR, la aplicación lo reconoce y despliega una foto del autor y un mensaje que invita al usuario a continuar con el recorrido. Internamente la aplicación guarda la información en la que está el usuario y la utiliza en la siguiente etapa: reconocimiento de obra. El reconocimiento de la obra se da una vez que el usuario está frente a la misma y toma una foto de ella que es procesada y en pocos segundos la aplicación responde con la imagen original de la obra y el usuario puede comenzar la interacción con la obra, a través de audioguías o realidad aumentada. Ver Figura 11.1

De esta manera es que se da el flujo de la aplicación a nivel de usuario, para llegar a un determinado cuadro de interés y así entonces interactuar con él. Pero este flujo es necesario representarlo en una serie de clases e instancias y con cierta invocación de métodos que cumplan las reglas de Objective-C con las herramientas existentes de desarrollo que provee Xcode. Para tener una idea de cómo se mapea el flujo de la aplicación en el lenguaje de desarrollo, en la Figura 11.2, se presenta el *Storyboard* de la misma, que muestra la relación entre las distintas clases. Se recuerda al lector que el *Storyboard* es una herramienta de programación gráfica, que permite generar instancias de clases y vínculos entre las mismas en forma visual a la vez de ser una representación gráfica de la interfaz de usuario. A su vez, a la Figura 11.2 se le agregó un número identificador en cada *ViewController* para poder referenciarlos en la medida que sea necesario detallar determinados aspectos de las clases involucradas.

En las próximas subsecciones se explican algunas de las clases implementadas en la aplicación y

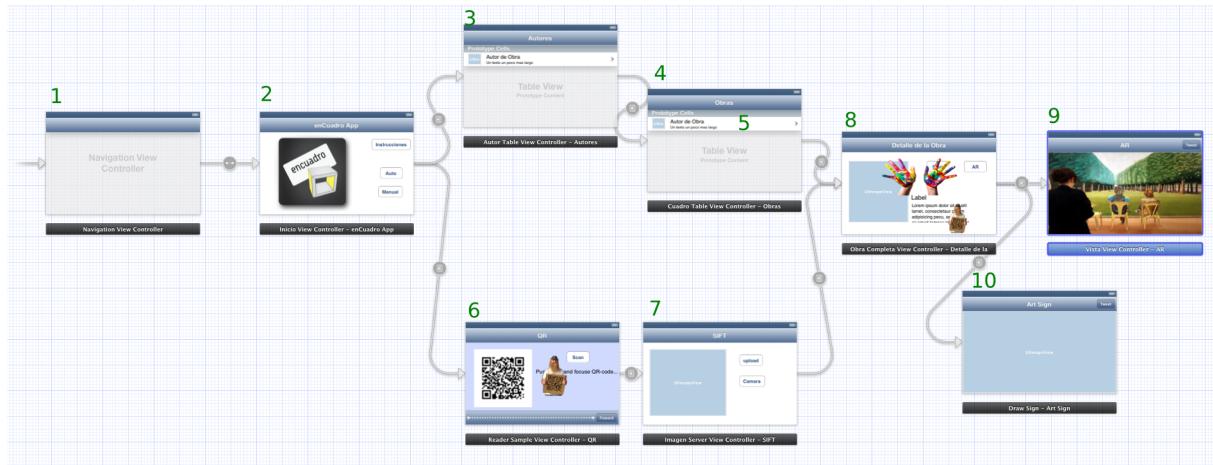


Figura 11.2: *Storyboard* de la aplicación

que además tienen cierta relevancia. Se muestran su rol dentro de la aplicación y sus principales características.

11.2.1. NavigationViewController

Esta clase se ve en la Figura 11.2, identificada con el número 1. La aplicación está embebida dentro de un *UINavigationController*. Esto implica que cada uno de los *ViewControllers* que tiene la aplicación es gestionado por esta clase. Es quien se encarga de la presentación y del pasaje de un *ViewController* a otro, creando y destruyendo instancias de cada uno. Está en esta clase la responsabilidad de manejar las jerarquías de los distintos *ViewControllers* así como de mantener cierta integridad visual utilizando las *Toolbars* ya sea arriba como encabezado o abajo al pie. Las *Toolbars* son botones que se pueden agregar en los extremos de los *ViewControllers* para realizar una funcionalidad específica.

El hecho de contar con una jerarquía permite entre otras cosas, la posibilidad de hacer un cambio (en la interfaz de usuario por ejemplo), en todos los *ViewControllers*, simplemente afectando a la clase *NavigationViewController* y sin necesidad de cambiar cada uno de ellos por separado. Esto resulta particularmente práctico en aplicaciones con bastantes *ViewControllers* y lo único que tiene que hacer el desarrollador es aclarar que ciertos atributos sean manejados por la clase encargada de la navegación dentro de la aplicación.

Por otra parte, es deseable tener un criterio común para todos los *ViewControllers* en la orientación de la aplicación con respecto a la orientación del dispositivo. Es decir, es posible lograr por ejemplo, que frente a rotaciones del dispositivo, la interfaz de usuario acompañe la rotación y gire también, o también es posible permitir que rotaciones del dispositivo en determinado sentido se vean reflejados en una rotación de la interfaz de usuario y otras no. Para esto se definen cuatro posibles posiciones para el dispositivo con ayuda del acelerómetro: *Portrait*, *Upside Down*, *Landscape Left* y *Landscape Right*. Las mismas se pueden ver en la Figura 11.3.

Para el caso particular de esta aplicación se optó por reimplementar la clase *UINavigationController* bajo el nombre *NavigationViewController* ya que se buscaba tener cierto control sobre las rotaciones de la interfaz de usuario, por lo que se decidió afectar los métodos que estuvieran a cargo de las rotaciones de interfaz de usuario. En particular se reimplementaron los métodos *supportedInterfaceOrientations* y *preferredInterfaceOrientationForPresentation* de la siguiente manera

```
- (NSUInteger)supportedInterfaceOrientations
```

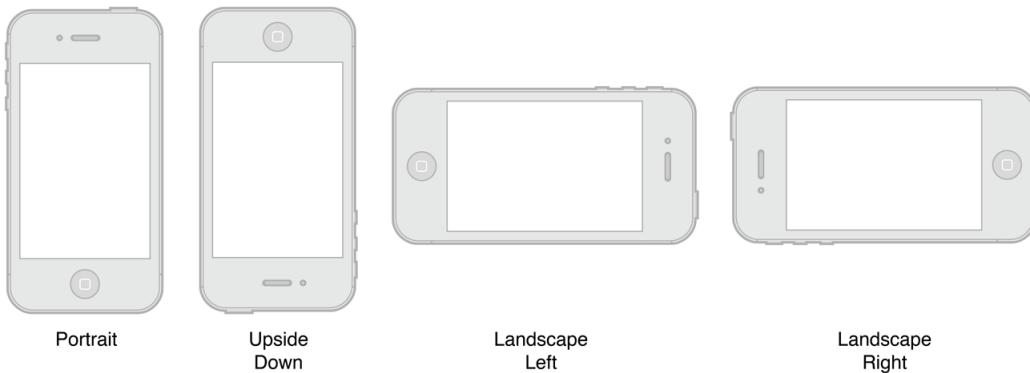


Figura 11.3: Orientaciones posibles del dispositivo.

```
{
    NSLog(@"supportedInterfaceOrientations NAVIGATION");
    return UIInterfaceOrientationMaskLandscapeRight;
}

- (UIInterfaceOrientation)preferredInterfaceOrientationForPresentation
{
    NSLog(@"preferredInterfaceOrientationForPresentation NAVIGATION");
    return UIInterfaceOrientationLandscapeRight;
}
```

Esto lo que hace es fijar la orientación de la interfaz de usuario a modo *LandscapeRight*. También hubiera sido posible lograrlo editando el archivo *Info.plist* que toda aplicación de Xcode tiene, agregando el ítem *SupportedInterfaceOrientations* y completando las opciones que se desean. Las rotaciones de interfaz de usuario son algo con bastante relevancia en las aplicaciones. En particular se optó por bloquear las rotaciones de interfaz, dejándola fija, para facilitar la reproyección de la realidad aumentada. De no haberlo hecho de esta manera, con cada rotación de la interfaz se tendrían que intercambiar los ejes de coordenadas en función del sentido de la rotación. Esto es posible de hacer ya que con cada rotación se ejecuta una serie de métodos en forma automática entre los cuales se encuentra el siguiente:

```
- (void) willRotateToInterfaceOrientation:(UIInterfaceOrientation)
toInterfaceOrientation duration:(NSTimeInterval)duration;
```

Dentro de dicho método sería posible hacer el ajuste de coordenadas correspondiente. La serie de métodos que son ejecutados al haber un evento del tipo rotación es algo que ha sufrido cambios recientes con la actualización de *software* a iOS 6.

11.2.2. InicioViewController

Este *ViewController* es la pantalla de inicio de la aplicación, identificado con el número 2 en la Figura 11.2. En la misma hay un botón que al ser presionado comienza un audio con instrucciones y una presentación sobre cómo es el recorrido y las funcionalidades con las que cuenta la aplicación. También hay dos botones más que dan al usuario la opción de elegir la forma de recorrer el museo: autónoma o automática. El botón de recorrido automático instancia al *ReaderSampleViewController* y el de recorrido autónomo instancia al *AutorTableViewController*.

11.2.3. UITableViewControllers

Para el recorrido manual, el usuario es el encargado de seleccionar el autor, luego las obras disponibles del autor seleccionado y luego se muestra un detalle de la obra seleccionada por el usuario presentando una instancia del *ViewController* llamado *ObraCompletaViewController*. Este recorrido que parece bastante intuitivo aparece en muchas aplicaciones de iOS en las que existen listas de datos. Un ejemplo son las aplicaciones que gestionan contenido musical que está ordenado en base a autores, dentro de los mismos, sus discos y dentro de los discos sus canciones. Como navegar en listas de datos es algo bastante frecuente, Xcode ya tiene implementada una clase llamada *UITableViewController*. En la Figura 11.4 se puede ver un ejemplo con varios tipos de tablas que organizan la información. Como se puede ver la tabla es una forma sencilla de organizar la información en la que existe una sola columna y muchas filas, llamadas celdas. También pueden existir secciones, con un encabezado y pie de sección. Volviendo a la aplicación lo que se hizo entonces

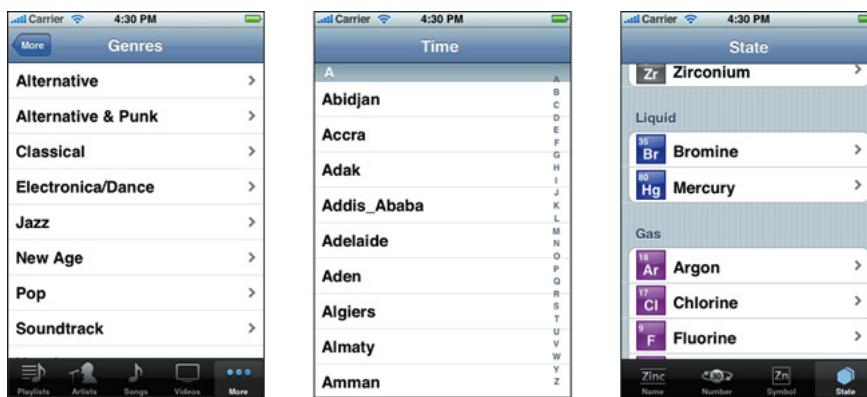


Figura 11.4: Ejemplos de TableViewControllers con distintos tipos de tablas.

fue crear varias clases que heredan de *UITableViewController* y manejar los contenidos de manera jerárquica. A continuación siguen dos clases que se resolvieron de esta manera.

11.2.3.1. AutorTableViewController

Esta clase (identificada con el número 3 en la Figura 11.2) hereda de *UITableViewController* y cumple la función de almacenar la lista de autores disponibles dentro del museo que a los efectos del prototipo como se dijo son: Figari, Blanes y Torres García. En lo que sigue se explican algunos detalles importantes que se tuvieron que comprender para poder organizar la información en tablas de datos (lo cual también aplica para la clase *CuadroTableViewController* que se describe en la siguiente subsección).

Uno de los métodos implementados por esta clase es el siguiente:

- (NSInteger)numberOfSectionsInTableView:(UITableView *)tableView

que por defecto retorna un 0. El mismo indica la cantidad de secciones con las que cuenta una tabla. Para que tenga sentido y al instanciarse la clase se vea algo de contenido tiene que retornar algo distinto de 0. Otro método importante es:

- (NSInteger)tableView:(UITableView *)tableView numberOfRowsInSection:(NSInteger)section

El mismo es el encargado de devolver un número con la cantidad de filas con las que cuenta la sección de la tabla. En esta implementación se devuelve la cantidad de autores.

Un tercer método, de mayor importancia, es el siguiente:

```
- (UITableViewCell *)tableView:(UITableView *)tableView cellForRowAtIndexPath:(NSIndexPath *)indexPath
```

El mismo es el encargado de devolver una *UITableViewCell* que es la que se despliega. Es en este método que se configura el formato de la celda. Para el caso de la aplicación se resolvió generar una clase que hereda de *UITableViewCell* que se llama *CuadroTableViewCell* y que tiene ciertas características como una imagen, autor y obra que son mostradas en la celda. En este método se asocian las características mencionadas de la celda en función del número de fila. Esta clase implementa un método *prepareForSegue* que le asigna un valor a la variable *opcionAutor* en función del autor seleccionado. Esto permite luego en la clase *CuadroTableViewController* desplegar distintas listas de cuadros en función del autor seleccionado.

11.2.3.2. CuadroTableViewController

Esta clase (identificada con el número 4 en la Figura 11.2) es muy similar a la clase *AutorTableViewController* recién descrita pero que difiere simplemente en su contenido. Los conceptos utilizados y métodos implementados son básicamente los mismos pero su contenido es un listado de obras en lugar de autores. Una especificación extra es que al instanciarse la clase se completa una lista de cuadros diferente en función del autor seleccionado. Así como en la clase *AutorTableViewController* en esta también se implementa el método *prepareForSegue* para poder completar los datos de la instancia de la clase con la que se está conectando, con los datos de la obra seleccionada (autor, obra, imagen, descripción, audio, ARid). El ARid es un identificador de realidad aumentada que asocia una realidad aumentada a cada cuadro.

11.2.3.3. CuadroTableViewCell

Esta es una clase sencilla que hereda de la clase *UITableViewCell* (identificada con el número 5 en la Figura 11.2) y simplemente tiene tres atributos asociados a nivel de interfaz de usuario: una imagen, un nombre de autor y un nombre de obra para cada celda de la tabla que se despliega.

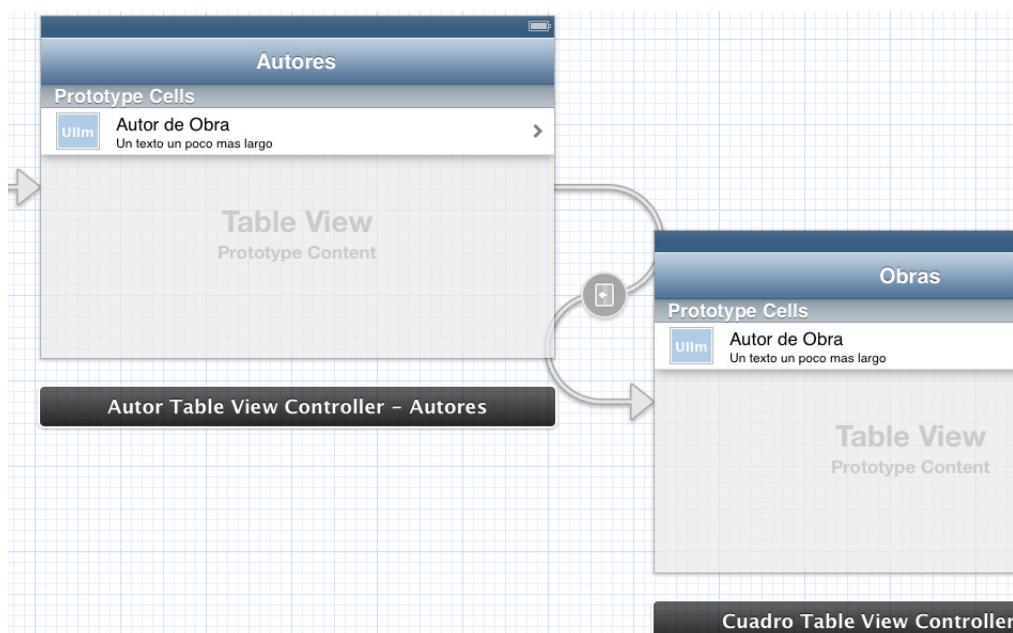


Figura 11.5: Autor y Cuadro TableViewControllers

11.2.4. ReaderSampleViewController

Este *ViewController* (identificado con el número 6 en la Figura 11.2) es el encargado de hacer la lectura de los códigos QR y de invocar los métodos necesarios para realizar la búsqueda de la zona del museo en la que se encuentra el usuario. Esto es, existe un código QR asociado a cada autor (Blanes, Figari y Torres García) y en base al código QR leído se despliega un texto y una imagen asociados al mismo. El funcionamiento de la decodificación se explica un poco más en detalle en la sección 11.3.

11.2.5. ImagenServerViewController

Este *ViewController* (identificado con el número 7 en la Figura 11.2) es el encargado de la comunicación con el servidor. Al instanciarse esta clase, también se instancia la clase *UIImagePickerController*, encargada de implementar una captura de imagen. Una vez que se toma una fotografía a la obra, la misma se muestra en una *UIImageView* y existen dos botones: uno de ellos simplemente dispara una nueva instancia del *UIImagePickerController* dando la opción de volver a tomar la fotografía y el otro botón inicia la comunicación con el servidor. Ver Figura 11.6.



Figura 11.6: Ejemplo de captura para reconocimiento SIFT

El botón encargado de la comunicación con el servidor, botón de *upload*, es un *segue* hacia el *ObraCompletaViewController*. Dentro del método *prepareForSegue*, encargado de preparar todo previo a la invocación de *ObraCompletaViewController* se invoca el método *uploadImage*. Este método genera un mensaje HTTP del tipo POST y se lo envía a la IP del servidor. En el cuerpo del mensaje se adjunta la foto tomada previamente y se le agrega una variable llamada *room*. Esta variable es completada previamente en el *ReaderSampleViewController* en base al QR detectado, dando información respecto de en qué sala/región del museo se encuentra el usuario (sala Figari, sala Blanes o sala Torres García). Esta variable lo que permite es tener un identificador para poder realizar la búsqueda de la imagen tomada en una base de datos más pequeña, que contenga solamente los cuadros de la región del museo en cuestión. En caso que el usuario se haya salteado la detección QR y haya seleccionado directamente la opción de tomar una fotografía a la obra para comenzar la comunicación con el servidor, entonces la variable *room* estará vacía y la búsqueda de la obra se realiza en toda la base de datos del museo. El gran valor agregado de la detección QR es la velocidad con la que el servidor devuelve información respecto de a qué obra se fotografió. Para la búsqueda con detección QR, los tiempos son claramente mejores (del orden de 3s en una LAN),

mientras que cuando el usuario se ahorra este paso, los tiempos aumentan al doble (del orden de 6s en una LAN).

Luego de establecida la conexión y enviada la consulta POST, el servidor responde con otra variable llamada *returnString*. Esta variable contiene un identificador de obra que indica qué obra fue fotografiada. Esto se logra mediante un archivo *upload.php* en el servidor que recibe la imagen y le ejecuta un algoritmo de detección de características llamado SIFT, que le retorna al PHP el identificador en cuestión. Detalles sobre el algoritmo SIFT se pueden ver más adelante en la sección 11.5. El archivo *upload.php* entrega esta información a la aplicación. La variable *returnString* es recibida por la aplicación con cierta nomenclatura en particular, que sigue la lógica Autor-Número, por ejemplo “Figari3” se corresponde con la obra número 3 de la base de datos del autor Figari. Con este identificador de obra, la aplicación le pide al servidor cierta información de interés acerca de la misma, como por ejemplo el nombre completo de la obra, el nombre de su autor y una breve descripción. El servidor cuenta con varias carpetas a las que la aplicación accede remotamente:

- (1) **autor:** contiene el nombre del autor de cada obra.
- (2) **obra:** contiene el nombre completo de cada obra.
- (3) **texto:** contiene una breve descripción de cada obra.
- (4) **imagen:** contiene una imagen de cada obra.
- (5) **audio:** contiene una audioguía asociada a cada obra.

Esta información solicitada es alojada en variables que son mostradas (imagenes, texto) y reproducidas (audio) en el siguiente *ViewController*, el *ObraCompletaViewController*. Ver Figura 11.7.

11.2.6. ObraCompletaViewController

Este *ViewController* (identificado con el número 8 en la Figura 11.2) simplemente es la presentación de la obra en la que se muestra una imagen del cuadro, título, autor, descripción y distintas opciones para interactuar con el mismo. Tiene dos botones y una animación que funciona como botón. Ver Figura 11.7. El primero de los botones dispara una audioguía relacionada con la obra que el usuario está contemplando. El otro botón conecta con el *VistaViewController*, encargado de mostrar la realidad aumentada, explicado en la sección 11.2.7. La animación que aparece funciona como *segue* hacia otro *ViewController*, llamado *DrawSign* que se explica más adelante en la sección 11.2.8.

11.2.7. VistaViewController

Este *ViewController* (identificado con el número 9 en la Figura 11.2) es el encargado de mostrar la realidad aumentada. Esta clase, al ser instanciada ejecuta el siguiente método:

```
- (void)viewWillAppear:(BOOL)animated
{
    NSLog(@"%@", @"VIEW WILL APPEAR VISTA");
    [super viewWillAppear:animated];
    [self hacerRender];
}
```



Figura 11.7: Pantalla con la obra completa

}

Este método se ejecuta justo antes de que el controlador despliegue el contenido de la pantalla, y como se ve, invoca al método homónimo de la clase superior y luego al método *hacerRender*, encargado de mostrar efectivamente la realidad aumentada. Antes de explicar los detalles de *hacerRender* se comentan algunos detalles generales de las aplicaciones iOS.

Como en cualquier programa, en las aplicaciones de Xcode, lo que se ejecuta al comenzar es el *main*. En este tipo de aplicaciones en particular, el *main* crea una instancia de la clase *appDelegate* (delegado de la aplicación). A su vez, al instanciarse al *appDelegate* se ejecuta el método *applicationDidFinishLaunching*. En este método, típicamente el código por defecto está vacío, pero cuando se trabaja con ISGL3D, este método crea un objeto de la clase *Isgl3dViewController* que hereda de *UIViewController*. Es sobre la instancia de *Isgl3dViewController* que se despliegan los *renders*. Aclarados estos puntos se pasa ahora a explicar lo que se hace en el método *hacerRender*. A continuación se muestran algunas de las partes más importantes del método:

```
app0100AppDelegate *appDelegate = (app0100AppDelegate *)[[UIApplication sharedApplication] delegate];
self.viewController=(Isgl3dViewController*)appDelegate.viewController;
```

Con lo anterior lo que se hace es generar una instancia de la clase *app0100AppDelegate* que es puntero al *appDelegate* de la aplicación. Luego, en la segunda línea se le asigna a la propiedad de la clase *VistaViewController* llamada *viewController* (que es de tipo *Isgl3dViewController*) la propiedad de igual nombre pero del *appDelegate* de la aplicación (que fue instanciada en el método *applicationDidFinishLaunching*). Luego se agregan las *views* *viewController:view* y *viewController:videoView* con valor de transparencia *alpha* nulo y se inicia una animación generando un efecto de *fade out* de la imagen y *fade in* del *render*. Este tipo de animaciones son sencillas de ejecutar con el framework Core Animation y permiten agregar efectos interesantes a cualquier *UIView*.

11.2.8. DrawSign

Esta clase (identificada con el número 10 en la Figura 11.2) hereda de *UIViewController* y está pensada para que el usuario pueda dibujar al tocar la pantalla. Un ejemplo de cómo queda el dibujo



Figura 11.8: Ejemplo de dibujo libre

se puede ver en la Figura 11.8. Se implementó haciendo una reimplementación de los siguientes tres métodos:

- (void)touchesBegan:(NSSet *)touches withEvent:(UIEvent *)event;
- (void)touchesMoved:(NSSet *)touches withEvent:(UIEvent *)event;
- (void)touchesEnded:(NSSet *)touches withEvent:(UIEvent *)event;

Cada vez que una instancia de una clase que hereda de *UIViewController* detecta un toque sobre la pantalla (evento *touch*), se invocan los métodos mencionados. La secuencia de invocaciones se da al comenzar el toque en la pantalla (*touchesBegan*), al desplazar el dedo sin levantarla de la pantalla (*touchesMoved*) y al finalizar el *gesture* levantando el dedo de la pantalla (*touchesEnded*). Un *gesture* es una forma característica de tocar la pantalla. Ejemplos de *gestures* existentes son: *touch*, *double touch*, *multi touch* entre otros. Se obtienen entonces las coordenadas del punto de toque sobre la pantalla invocando el siguiente método:

```
[touch locationInView:self.view]
```

donde *touch* es del tipo *UITouch* y tiene propiedades que dependen del evento. Una vez que se tienen las coordenadas del punto de contacto en la pantalla se guarda esta posición y al obtener una nueva posición en la pantalla (luego de desplazar el dedo en *touchesMoved*), se dibuja una línea entre el punto actual y el anterior con el método siguiente:

```
[image.image drawInRect:CGRectMake(0, 0, self.view.frame.size.width,
                                     self.view.frame.size.height)];
```

El método *drawInRect* sirve para dibujar en forma 2D sobre *UIViews* y fue utilizado extensivamente en este proyecto. Finalmente en el método *touchesEnded* lo que se hace es dibujar una línea en el punto actual y sí mismo, generando un punto final al levantar el dedo de la pantalla. Otra característica interesante a mencionar respecto de la capacidad de responder a eventos *touch* es el reconocimiento de *gestures* que pueden ser nativos o incluso creados por el propio desarrollador. De esta manera si se quiere reconocer un *double touch* por ejemplo, se puede invocar el siguiente método:

```
[touch tapCount];
```

que devuelve la cantidad de veces que se tocó la pantalla en un intervalo corto de tiempo. Esto fue utilizado en esta clase para borrar lo dibujado y poder comenzar a dibujar nuevamente.

A esta clase también se le agregó una *IBAction* que genera un *tweet* con el dibujo generado por el usuario. El mismo es logrado generando una instancia de la clase *TWTweetComposeViewController* y agregándole un texto e imagen con los siguientes métodos:

```
[controller setInitialText:text];
[controller addImage:img];
```

donde *text* e *img* son el texto del *tweet* y la imagen adjunta. Finalmente se presenta la *view* del *TWTweetComposeViewController* y una vez finalizado se vuelve a la instancia *DrawSign*.

11.2.9. TouchVista

Esta clase hereda de la clase *UIView* y se creó para poder manejar eventos *touch* en *ViewControllers* que tienen varias *subviews* y que interesa que se dispare un evento al tocar tan sólo una de ellas en determinada área de la pantalla. Entonces lo que se hace en esos casos es agregar a la *subview* en cuestión una instancia de *TouchVista* en forma transparente por encima y del mismo tamaño. De esta manera al tocar la *subview* se toca en realidad la instancia de *TouchVista* y se invoca el método *touchesBegan*. Este método simplemente configura una bandera y configura lo siguiente:

```
[super touchesBegan:touches withEvent:event];
```

Lo que se hace en el código anterior es invocar al método *touchesBegan* de la clase superior. Para el caso en que se tiene un *ViewController*, con una *subview* del tipo *TouchVista* transparente, entonces esta línea invoca directamente el método *touchesBegan* del *ViewController*. Dos de los *ViewControllers* que utilizan esto son *VistaViewController* y *ObraCompletaViewController*.

11.2.10. Realidad Aumentada en ISGL3D

Para realizar realidad aumentada se necesita poder hacer un *render* por encima de las imágenes capturadas por la cámara del dispositivo en tiempo real. Sin embargo, cuando se crea un proyecto de ISGL3D, este permite realizar *renders* pero sobre un fondo estático y gris (o cualquier otro color configurable). Resulta entonces necesario configurar el proyecto de manera de reemplazar al fondo antes mencionado por imágenes capturadas por la cámara. Para lograr esto, hubo que trabajar sobre las clases *Isgl3dViewController* y *app0100AppDelegate*. A continuación se muestran algunas modificaciones sobre estas dos clases

```
UIImageView* vistaImg = [[UIImageView alloc] init];

/* Se ajusta la pantalla*/
UIScreen *screen = [UIScreen mainScreen];
CGRect fullScreenRect = screen.bounds;

[vistaImg setCenter:CGPointMake(fullScreenRect.size.width/2, fullScreenRect.size.height
[vistaImg setBounds:fullScreenRect];

[self.window addSubview:vistaImg];
[self.window sendSubviewToBack:vistaImg];
 viewController.videoView = vistaImg;
```

Con esto se ajusta el atributo *videoView* de la propiedad *viewController* que pertenece a la clase *app0100AppDelegate* y es instancia de la clase *Isgl3dViewController*.

Como fue mencionado en el capítulo 1, en Xcode, si se quiere simplemente hacer una filmación, sacar fotos o acceder a la galería de las fotos, se utiliza generalmente instancias de la clase *UIImagePickerController*. Esta última clase se instancia en la aplicación en otras clases, como ser *ImagenServerViewController*. Si lo que se desea es acceder a los píxeles de las imágenes capturadas, para luego poder procesarlos en tiempo real, entonces la forma más indicada es usando el conocido framework *AVFoundation*. En la clase *Isgl3dViewController*, en el método *viewDidLoad* está toda la configuración necesaria para la utilización de *AVFoundation*. A continuación se muestra el código con sus comentarios sobre esta configuración.

```
/*Creamos y seteamos la captureSession*/
self.session = [[AVCaptureSession alloc] init];
self.session.sessionPreset = AVCaptureSessionPresetMedium;

/*Creamos al videoDevice*/
self.videoDevice = [AVCaptureDevice defaultDeviceWithMediaType:AVMediaTypeVideo];

/*Creamos al videoInput*/
self.videoInput = [AVCaptureDeviceInput deviceInputWithDevice:self.videoDevice error:&error];

/*Creamos y seteamos al frameOutput*/
self.frameOutput = [[AVCaptureVideoDataOutput alloc] init];

self.frameOutput.videoSettings = [NSDictionary dictionaryWithObject:[NSNumber numberWithInt:kCVPixelFormatType_32BGRA] forKey:kAVVideoMinFrameDurationKey];

/*Ahora conectamos todos los objetos*/
/*Primero le agregamos a la sesión el videoInput y el videoOutput*/

[self.session addInput: self.videoInput];
[self.session addOutput: self.frameOutput];
```

Como se ve, es necesario crear una sesión de captura, luego un dispositivo de captura y una salida de los datos y agregarlos a la sesión. También se puede configurar el tipo de captura de la cámara (tiene que ser soportado por el *hardware*, sino se genera un error en este punto). Otra cosa importante que se hace en la clase *Isgl3dViewController* es la configuración del *multi-threading*. A continuación se muestra el código que logra esto.

```
dispatch_queue_t processQueue = dispatch_queue_create("procesador", NULL);
[self.frameOutput setSampleBufferDelegate:self queue:processQueue];
dispatch_release(processQueue);
```

Con esto lo que se hace es hacer una instancia de una *Queue*, que representa una cola de procesamiento. De esta manera se puede hacer que ciertas tareas se alojen en esa instancia de cola, que lógicamente es otra distinta que la cola de procesamiento principal (*mainQueue*). Esto mismo es lo que se hace en la segunda línea del código anterior, diciendo que el *Delegate* de los datos de salida (*frameOutput*) es la propia clase y que ese *Delegate* se ejecute en la *Queue* que se instanció en la línea anterior. De esta manera todo lo que sea invocado por el *Delegate* en forma periódica será enviado a una cola distinta de la principal, pudiendo tener entonces, una cola de procesamiento separada de la cola de interfaz de usuario. Esto es algo ampliamente utilizado y es una recomendación

de la documentación de iOS, pues se basa en los conceptos de tener la mayor atención posible a la interfaz de usuario, impidiendo en lo posible dejar al usuario esperando por algún eventual procesamiento que se esté llevando a cabo.

Finalmente se da comienzo a la sesión:

```
[self.session startRunning];
```

Como la clase *Isgl3dViewController* implementa el protocolo *AVCaptureVideoDataOutputSampleBufferDelegate*, una vez que comienza la sesión, se invoca cuadro a cuadro el siguiente método:

```
-(void) captureOutput:(AVCaptureOutput *)captureOutput didOutputSampleBuffer:(CMSampleBufferRef)sampleBuffer fromConnection:(AVCaptureConnection *)connection;
```

donde *sampleBuffer* es una referencia al *buffer* que contiene los píxeles de la cámara en ese momento. Así entonces, se accede a los píxeles y se invoca dentro del *captureOutput* periódicamente al método *procesamiento*, encargado de procesar la imagen recibida por la cámara.

11.3. QR

11.3.1. Identificadores QR. Una realidad

El uso de los identificadores QR (Quick Response), es cada vez más generalizado. Últimamente, debido al incremento significativo del uso de *smart devices*, el hecho de poder contar con una cámara, cierto poder de procesamiento y por lo general hasta una conexión móvil a internet, hace que sea cada vez más frecuente encontrar aplicaciones con el poder de reconocer QRs. Comenzaron utilizándose en la industria automovilística japonesa como una solución para el trazado en la línea de producción, pero su campo de aplicación se ha diversificado y hoy en día se pueden encontrar también como identificadores de entradas deportivas, tickets de avión, localización geográfica, vínculos a páginas web y en algunos casos también como tarjetas personales.

11.3.2. ¿Qué son realmente los QR?

Se puede decir que los QRs tienen muchos puntos en común con los códigos de barras pero con la ventaja de poder almacenar mucho más información debido a su bidimensionalidad. Existen distintos tipos de QR, con distintas capacidades de almacenamiento que dependen de la versión, el tipo de datos almacenados y del tipo de corrección de errores. En su versión 40 con detección de errores de nivel L, se pueden almacenar alrededor de 4300 caracteres alfanuméricos o 7000 dígitos (frente a los 20-30 dígitos del código de barras) lo cual lo hace muy flexible para cualquier tipo de aplicación de identificación.

En la Figura 11.9 se pueden ver las distintas partes que componen un QR, como por ejemplo el bloque de control, compuesto por las tres esquinas idénticas que dan información de la posición, la información de alineamiento y el patrón de sincronismo; así como también la indicación de versión, formato y la corrección de errores. Fuera de toda esa información, que podría verse como el encabezado, haciendo analogía con los paquetes de las redes de datos, se encuentran los datos propiamente dicho, que podrían verse como el cuerpo del paquete.

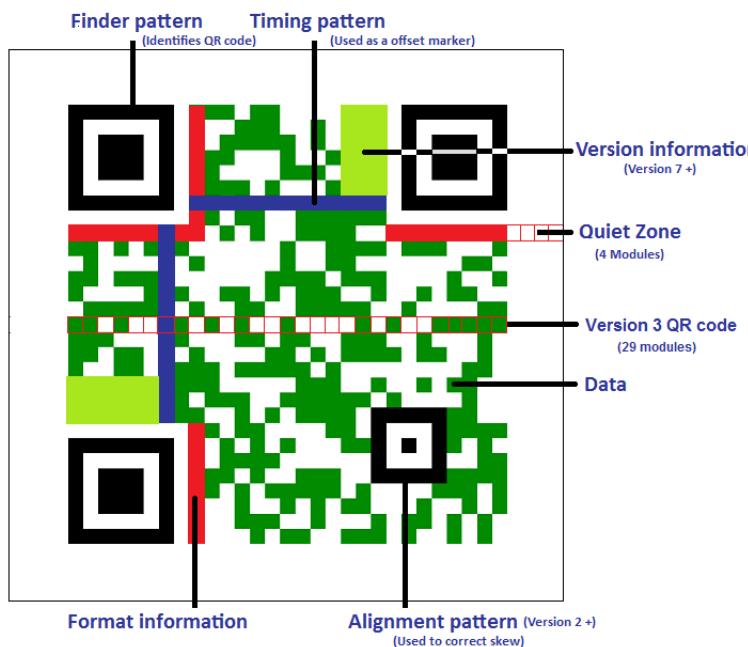


Figura 11.9: Las distintas componentes de un QR. Fuente (poner fuente).

11.3.3. Codificación y decodificación de códigos QR

Es fácil darse cuenta que la codificación resulta mucho más sencilla que la decodificación. Para la codificación es necesario comprender el protocolo, las distintas variantes y el tipo de información que se pretende almacenar. Sin embargo, para la decodificación, además de tener que cumplir con lo anterior, es necesario contar con buenos sensores y ciertas condiciones de luminosidad y distancia que favorezcan a la cámara y se traduzcan en buenos resultados luego de la detección de errores. Si bien la plataforma es importante para lograr buenos resultados, dada una plataforma, existen variadas aplicaciones tanto para iOS como para Android que cuentan con performances bastante diferentes en función del algoritmo de procesamiento utilizado.

Debido a que el centro del presente proyecto no fue la codificación y decodificación de QRs, y que además ya existen distintas librerías que resuelven muy bien este problema, se optó por investigar varias de ellas e incorporar la más adecuada a la aplicación.

Entre todas las librerías que resuelven la decodificación, las llamadas ZXing y ZBar son quizás las más destacadas, por su popularidad, simplicidad y buena documentación para la fácil implementación. ZXing, denominada así por “Zebra Crossing”, es una librería gratis y en código abierto desarrollada en java y que tiene implementaciones que están adaptadas para otros lenguajes como C++, Objective-C y JRuby, entre otros.

Por su parte ZBar también tiene soporte sobre varios lenguajes y cuenta con un kit de desarrollo interesante para lograr fácilmente aplicaciones que integren el lector de QR. Se trabajó sobre el código de ejemplo que contiene la implementación de las clases principales para obtener un lector y finalmente se optó por utilizar esta librería para los fines de la aplicación. El lector del código de ejemplo consta de una clase *ReaderSampleViewController* que hereda de *UIViewController* y que implementa un protocolo llamado *ZBarReaderDelegate*. Al presionarse el botón de detección se crea una instancia de la clase *ReaderSampleViewController* y se presenta la vista previa de la cámara. Luego el protocolo se encarga de la captura y procesamiento del QR almacenando como re-

sultado la información embebida en este en la variable denominada *ZBarReaderControllerResults*. Esta variable luego se mapea en una *hash table* con el contenido en formato *NSDictionary*. De esta manera se accede fácilmente al contenido en formato legible y es fácil de hacer una lógica de comparación y búsqueda en una base de datos.

11.3.4. El QR en la aplicación

Para el caso particular de la aplicación se optó por tener un identificador QR para cada uno de los tres artistas elegidos del Museo Nacional de Artes Visuales (MNAV). De esta manera, para el caso del recorrido del museo de forma automática, es posible determinar la posición del usuario utilizando imágenes QR debidamente ubicadas en cada zona. Esto sirve como localización y también sirve para lograr que el paso siguiente, que es la identificación de la obra que el usuario tiene enfrente, sea mediante una búsqueda en una base de datos discriminada por autor como se explicó en la sección 11.6. Es decir, si el usuario no escanea el QR la búsqueda de la obra a identificar se hará en una base de datos global del museo, pero en el caso que el usuario sí decida escanear el QR, entonces se cuenta con la posibilidad de realizar la búsqueda en una base de datos reducida y por lo tanto más veloz.

11.3.5. Buen gusto para los QR

La opción de usar los QR de una manera distinta ha comenzado a ser notoria en los últimos tiempos. Hay quienes desafían a la información *cruda de 1s y 0s* incorporando imágenes y modificando colores y contornos en los QR tradicionales para lograr un valor estético además del funcional. Véase en la figura 11.10 un ejemplo de cómo puede lograrse el mismo resultado pero con el valor agregado de originalidad.



Figura 11.10: Ejemplo de un QR creativo. Fuente (poner fuente).

11.4. Servidor

Si bien el desarrollo de la aplicación busca lograr un prototipo y no una aplicación comercial, por lo que la cantidad de imágenes y datos en general es pequeña y por lo tanto puede ser almacenada dentro de la aplicación, por prolijidad y escalabilidad resulta imprescindible contar con un servidor. Es necesario un servidor que guarde la toda la información y a la vez realice algo de procesamiento, siempre y cuando este no deba hacerse en tiempo real. Se decidió entonces, almacenar toda aquella información relevante en cuanto a registro de obras (imágenes, títulos, autores y descripciones), audioguías, videos, modelos y animaciones utilizadas para la realidad aumentada en un servidor que hubo que implementar. El servidor debe estar ubicado dentro del museo y se debe poder interactuar con él por medio de la LAN (54 Mbps). Si bien es cierto que el servidor podría perfectamente ser remoto, con acceso por medio de internet, su desempeño a nivel de tiempos bajaría notoriamente.

11.4.1. Creando el servidor

Para la creación del servidor se buscó primeramente la alternativa de hacerlo sobre una máquina con sistema operativo con núcleo Linux, distribución Ubuntu, ya que se creyó sería más sencillo. Luego de realizada esta tarea, se estudió la posibilidad de tener el servidor corriendo sobre una plataforma Mac OS X y de manera sorpresiva, en este segundo caso, el objetivo se logró de forma mucho más sencilla. A continuación se explica paso por paso como se implementaron los servidores en uno y otro sistema operativo.

11.4.1.1. Servidor LAMP

Se le llama LAMP a la combinación de todas las herramientas necesarias para lograr un servidor web: Linux (Sistema Operativo), Apache (Servidor Web), MySQL (Gestor de base de datos) y PHP/Perl/Python (lenguaje de programación del lado del servidor). Se comenzó entonces por instalar un servidor web Apache, que tiene como principales ventajas el hecho de ser multiplataforma, gratis y de código abierto. La descarga y la instalación son inmediatas, se abre un terminal y se escribe:

```
sudo apt-get install apache2
```

Una vez finalizada la instalación de este conjunto de paquetes ya se cuenta con el servidor y mediante los siguientes comandos, uno puede dar inicio y fin al mismo:

```
sudo /etc/init.d/apache2 start  
sudo /etc/init.d/apache2 stop
```

Luego de tener instalado Apache, se procede a instalar php:

```
sudo apt-get install php5
```

Para este proyecto en particular, no fue necesario instalar MySQL.

11.4.1.2. Servidor en Mac OS X

A diferencia de otros sistemas operativos, Mac OS X cuenta por defecto con Apache y Php, no así con MySQL. Para tener un servidor entonces, sólo resta activarlos y si además se quiere contar con MySQL, es necesario instalarlo. La activación del servidor es un proceso muy simple, del cual existe bastante material disponible en internet.

Una vez instalado el servidor en cualquiera de ambos sistemas operativos, se puede corroborar su correcto funcionamiento abriendo cualquier navegador web y digitando la dirección IP de la máquina en la que se instaló el servidor (si se está en ella misma, se puede digitar la dirección del bucle local). Se deberá ver la página que por defecto muestra el servidor, la de la figura 11.11.



Figura 11.11: Página por defecto del servidor Apache.

11.4.1.3. Aspectos a mejorar del Servidor

Como se mencionó el servidor se implementó sin MySQL y tampoco con ningún otro gestor de base de datos, lo que hace que cualquier modificación sobre la información de los cuadros, ya sea en cuanto a imágenes, descripciones, audioguías o cualquier información que se quiera modificar dentro del servidor haga que quien administre el mismo tenga la necesidad de comprender en cierta medida cómo está implementado y tener cierto dominio técnico. Para el caso de aplicar esto en un museo y pensando que quienes administren la información de los cuadros sea gente encargada del museo no especializada en aspectos de software, es deseable que la interfaz de gestión del servidor tenga un entorno más amigable. Esto es algo que se podría mejorar para un futuro en caso de querer continuar trabajando con el presente proyecto para darle más completitud y usabilidad.

11.5. SIFT

SIFT fue utilizado en el presente proyecto para el reconocimiento de las obras contempladas por los usuarios, cuando estos deciden realizar el recorrido interactivo del museo de forma automática. El algoritmo corre en el servidor y el código utilizado es una adaptación propia de la implementación en C que está en la librería VL-Feat. VL-Feat es una librería gratis y en código abierto que implementa algoritmos populares de visión artificial que puede ser descargada de su página web oficial [?].

Se tiene entonces, para cada obra en cuestión, una lista de 128 descriptores invariantes a factores de escala, traslación, rotación y parcialmente invariantes a cambios de iluminación y afinidades; almacenada en el servidor. Cuando el usuario se encuentra frente a una obra determinada, este le toma

una fotografía y esta es subida al servidor, en donde es procesada con SIFT y luego sus descriptores son comparados contra todos los descriptores de la base de datos, o al menos los correspondientes a la región del museo en donde el usuario se encuentra. La obra con más descriptores en común con la imagen en cuestión será la que el usuario contempla.

11.6. Comentarios finales sobre la implementación

Hasta el momento se mencionaron una cantidad de herramientas más que interesantes, que reunidas logran un recorrido interactivo para uno o más museos. Sin embargo, no se expusieron cuáles son las aplicaciones puntuales de realidad aumentada que se dijo se iba a hacer sobre las obras. Por otra parte, aunque dejando afuera aspectos estéticos y artísticos, en el capítulo 10 se mostró como se resolvieron los aspectos técnicos necesarios para llevar a cabo una aplicación final real de realidad aumentada. Así entonces el generar un *render* de un perro junto a un sillón (caso de uso “modelos”), que claramente puede llegar a ser de muy poco interés para un museo, resuelve el mismo problema que si se quisiera generar un *render* con el modelo de José Artigas. Asimismo ser capaz de responder frente a al toque en la pantalla del modelo de un cubo y así entonces actuar en consecuencia (caso de uso “interactivo”), resuelve el mismo problema que animar al modelo de José Artigas si este es tocado. Lo que se tiene en este caso (y efectivamente se logró), es una escultura digital e interactiva de Artigas, que sólo puede visualizarse a través de un *iPad*. Ver figura 11.12.

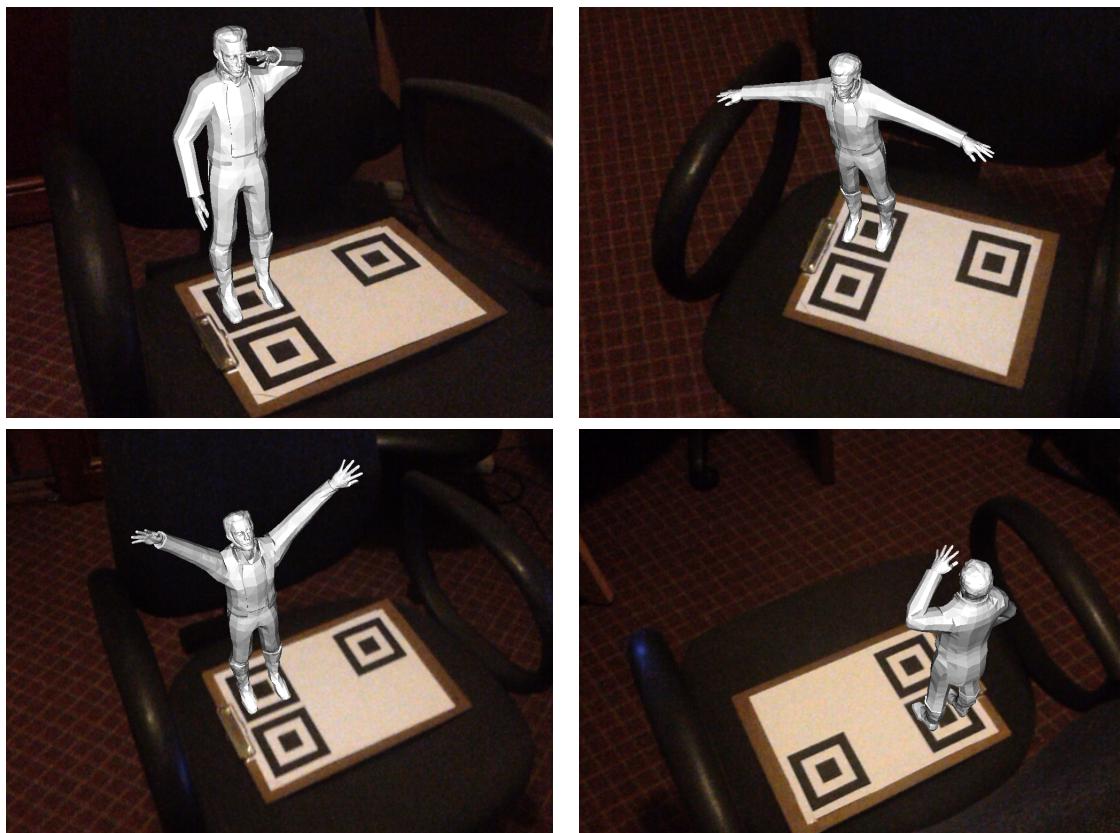


Figura 11.12: Escultura digital e interactiva de Artigas vista desde ángulos distintos y en posiciones distintas.

Lo mismo sucede con el caso de uso “video”, que proyecta un video sobre el marcador, que

puede facilmente adaptarse para proyectar un video de interés sobre una obra, parte de ella o incluso un mapa con información del museo o cualquier otra cosa. Así entonces logrando una audioguía interactiva, donde el usuario puede, mientras se le habla de la obra, ver fotos y videos relacionados a la misma. A la fecha se está buscando aplicar, dentro de lo posible, los desafíos técnicos resueltos en cada uno de los casos de uso mencionados en el capítulo 10 a diferentes implementaciones que puedan llegar a resultar atractivas para uno o varios museos. Las opciones son muchísimas y se cree que este proyecto deja una puerta abierta a seguir explorando ideas innovadoras.

Bibliografía

- [1] J. García Ocón. Autocalibración y sincronización de múltiples cámaras plz. 2007.
- [2] B. Furht. *The Handbook of Augmented Reality*. 2011.
- [3] C. Avellone and G. Capdehourat. Posicionamiento indoor con señales wifi. 2010.
- [4] Philip David, Daniel Dementhon, Ramani Duraiswami, and Hanan Samet. Simultaneous pose and correspondence determination using line features. pages 424–431, 2003.
- [5] Philip David, Daniel Dementhon, Ramani Duraiswami, and Hanan Samet. Softposit: Simultaneous pose and correspondence determination. pages 424–431, 2002.
- [6] Daniel F. DeMenthon and Larry S. Davis. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15:123–141, 1995.
- [7] R. Grompone von Gioi, J. Jakubowicz, J. M. Morel, and G. Randall. Lsd: A fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):722–732, April 2010.
- [8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.
- [9] V. Lepetit and P. Fua. Monocular model-based 3d tracking of rigid objects: A survey. *Foundations and Trends in Computer Graphics and Vision*, 1(1):1–89, 2005.