_ TELECOM ESCUELA TÉCNICA VLC SUPERIOR DE UPV INGENIEROS DE TELECOMUNICACIÓN

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

GRADO DE INGENIERÍA EN TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

RECONOCIMIENTO DE GESTOS MEDIANTE IMAGENES DE PROFUNDIDAD

Fiorella Anneth Jhonson García
Tutor : Antonio José Albiol Colomer

Trabajo Fin de Grado presentado en la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universitat Politècnica de València, para la obtención del Título de Graduado en Ingeniería de Tecnologías y Servicios de Telecomunicación
Curso 2016-17

A mis padres ,
por su incondicional apoyo.

RESUMEN

El objetivo del presente proyecto es la creación de un programa en lenguaje C++, que sea capaz de reconocer gestos realizados por un sujeto situado delante de una cámara.

Para ello, en primer lugar se han obtenido una secuencia de imágenes de distancias, mediante una cámara de profundidad, posteriormente estas imágenes son procesadas mediante una serie bloques en los que se ha divido el programa, cada uno de ellos arrojará un resultado, ya sea numérico o lógico, que será posteriormente utilizado por los siguientes bloques.

Los bloques en los que se ha divido el programa son tres, el primero detecta las manos del sujeto, el segundo detecta si ha habido movimiento (y por tanto se ha realizado un gesto), y el último detecta el tipo de gesto que se ha realizado.

Por otro lado, se pretende presentar al lector tres de las técnicas más destacadas para la adquisición de imagen 3D, que son la Estereovisión, la luz estructurada y tiempo de vuelo. Además de exponer algunas de las técnicas más empleadas en tratamiento de imagen, como lo son la Morfología y la Segmentación.

ABSTRACT

The main goal of this project is the creation of a program based on C++ language which is able to detect the gestures of a person placed in front of the camera.

The first step is recording the scene's depth images with a depth camera. Subsequently, these images will be processed in separate blocks of the program. Each one of these blocks will provide a result which can be either numerical or logical and which will be used later on in another block of the program.
The program has been divided into three blocks. The first one detects hands, the second one checks if the person has moved his hands, and the last one determines what kind of gesture has been performed.
Furthermore, we will also present three of the most important 3D imaging techniques, namely Stereovision, Structured Light and Time of Flight. Moreover, we will expound on some of the most used image processing techniques, such as Segmentation and Morphology.

Índice

1. INTRODUCCIÓN

En la actualidad cada vez se hace más presente la necesidad de una interfaz humano – máquina, por ello se han desarrollado diferentes algoritmos de detección de gestos.

La proyección de un mundo en 3D a uno en 2D conlleva la pérdida de información de la profundidad, lo que implica que, dada una imagen 2D, la geometría de la escena 3D observada no pueda ser reconstruida sin ambigüedad. La naturaleza ha satisfecho la necesidad de percepción de la profundidad para los seres humanos y la mayoría de los animales con un sistema de visión binocular, que basándose en la disparidad binocular extrae información a partir de un par de proyecciones 2D. Para el ser humano es una acción que no requiere esfuerzo alguno, sin embargo la reconstrucción de geometrías 3D mediante diferentes sistemas de sensores, se ha convertido en un reto tecnológico importante.
A lo largo de las pasadas décadas se han propuesto diferentes tecnologías con este fin, los últimos avances en óptica electrónica , diseño de sensores y potencia computacional han conseguido altas resoluciones (> 300 kpx) y un muestreo temporal lo más cercano posible al tiempo real (>=30Hz) [9] en la adquisición de imágenes 3D.

1.1.Objetivo

El objetivo del proyecto es la creación de un programa, que procese las imágenes obtenidas mediante una cámara profundidad, que utiliza la técnica de luz estructurada para obtener la profundidad de la escena, frente a la cual está situado un sujeto, para decidir si éste está o no realizando un gesto.

1.2.Estructura de la memoria

Tras haber expuesto el objetivo del proyecto, a continuación se presentan los capítulos en los que se ha dividido.
En primer lugar tenemos el capítulo 2, correspondiente a las técnicas de adquisición de imagen 3D.
El capítulo 3 describe el funcionamiento de una cámara de profundidad así como los elementos que la componen.
El capítulo 4 presenta dos técnicas de tratamiento de imagen, que son utilizadas en este proyecto, morfología y erosión.
En el capítulo 5 se describen los diferentes bloques de los que está compuesto el programa, así como la finalidad de cada uno.

2. Técnicas de adquisición de imagen 3D

A continuación se presentan las tres principales tecnologías de reconstrucción de superficies en 3D, la visión estereoscópica, luz estructurada y tiempo de vuelo.

2.1.Visión estereoscópica

La visión estereoscópica se fundamenta en el sistema visual humano, pues para ser capaces de percibir la profundidad se hace uso de dos imágenes, las que capta cada ojo correspondientemente, debido a la disposición horizontal de los ojos, cada una de estas imágenes es percibida con una perspectiva diferente, es decir con un ligero desplazamiento entre ambas, lo que hace posible la percepción de la profundidad.

Separación interocular y separación interaxial

La separación interocular técnicamente se refiere a la separación entre los centros de los globos oculares, la distancia típicamente aceptada es de alrededor de 65 mm para un hombre adulto[1]. La separación interaxial es la distancia que hay entre los lentes de las cámaras.

Figura 2.1. Separación interaxial en cámara estereoscópica [1]

Funcionamiento:

Como se ha explicado anteriormente para la percepción de la profundidad de una escena, son necesarias dos imágenes de la misma, por lo que son necesarias dos cámaras, obteniendo así dos perspectivas diferentes.

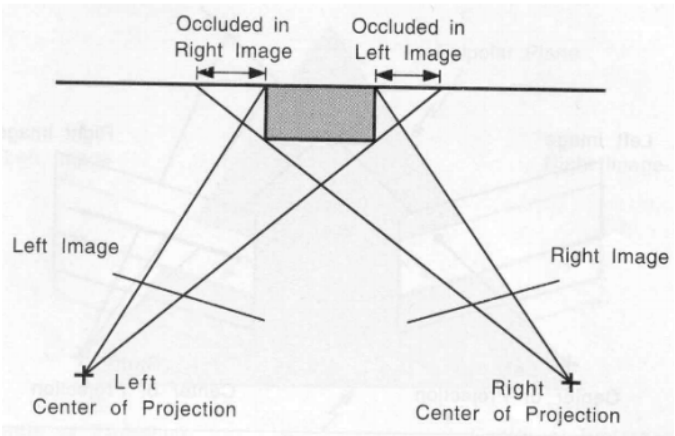

Figura 2.2. Colocación de las cámaras para la adquisición de escena por estéreo visión, puede verse como existen puntos que ven ambas cámaras y otros que solo son capturados por una de ella, pero el método es capaz de solventarlo [2]

Estéreo Visión

Para obtener información de profundidad de la escena mediante el método estereoscópico, es necesario, realizar una serie de cálculos sobre el par de imágenes obtenidas. De manera que dados dos píxeles que corresponden un mismo punto en la escena, se pueda, mediante el conocimiento de las condiciones de calibración del sistema, parámetros como la distancia focal y separación entre los centros de las cámaras, se pueda realizar el cómputo de la profundidad.

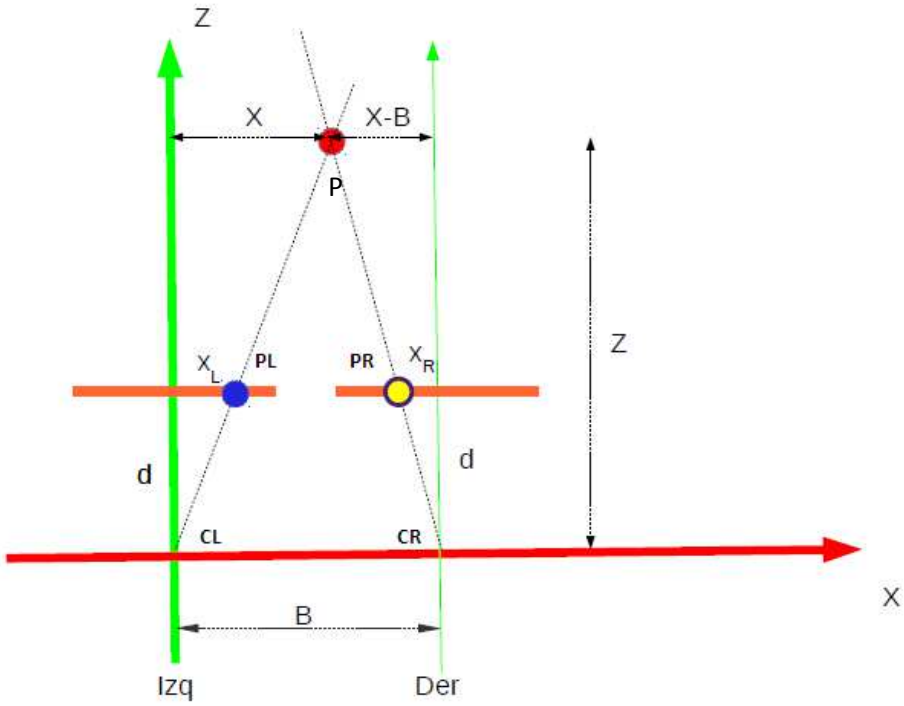

Figura 2.3. Imagen que muestra la disposición de las dos puntos sobre los planos imagen de cada una de las cámaras (azul y amarillo respectivamente), que genera punto de la escena (rojo), que es el punto de interés, cuyas coordenadas se desean conocer.

Triangulación

Proceso en el cual, dados 2 puntos en dos imágenes (uno en cada en imagen) que corresponden a un mismo punto en el espacio, del cual se quiere determinar su profundidad.
La situación planteada se refleja en la figura (2.3), de manera que el escenario en el que se basarán los cálculos, consta de los siguientes elementos:

- Dos puntos con coordenadas xr y xl, conocidas, misma coordenada z
- Punto desconocido P, coordenada X y Z desconocidas.
- Distancia focal : d (conocido)
- Distancia entre los centros de las cámaras B. (conocido)

Aplicando semejanza de triángulos, en la zona encerrada por los puntos CR , P y la línea paralela al eje z , obtenemos la siguiente ecuación:

$$\frac{X - B}{Z} = \frac{xR}{d} \qquad (2.1)$$

Y volviendo a aplicar el mismo teorema, esta vez en la zona formada por los puntos CL, P y eje z, se obtiene:

$$\frac{X}{Z} = \frac{xL}{d} \qquad (2.2)$$

Despejando X de la ecuación (2.2) y sustituyéndola en (2.1), obtenemos el valor de Z:

$$Z = \frac{Bd}{xR - xL} \qquad (2.3)$$

Hasta ahora se ha supuesto conocida la relación entre un par de puntos sobre el cuál se realiza la triangulación, es decir la correspondencia entre un punto de una imagen izquierda con otro en la derecha, a esto se le denomina búsqueda de correspondencias.

Problema de correspondencia

Define el hecho de encontrar la correspondencia entre un punto en una imagen y un punto en la otra , es decir, dado un punto en una de las imágenes que representa un punto en la escena, se debe encontrar en la otra imagen ese mismo punto de la escena.

Para facilitar esta búsqueda se hace uso de la geometría epipolar, de manera que en lugar de buscar el correspondiente a un punto perteneciente a una imagen, en la otra, no se realiza búsqueda en toda la imagen, si no en la línea epipolar correspondiente a dicho punto.

En la geometría presentada en la figura (2.4), se observan tres planos, 2 planos imagen que sería donde se formarían las dos imágenes de la escena, que son atravesados por un segmento (baseline) que une los puntos C y C', que son una representación de los respectivos centros de las cámaras y un tercero que es el llamado plano epipolar, que contiene el baseline, además de la línea epipolar que es la intersección del plano imagen con el plano epipolar.

Rotando el plano epipolar con diferentes ángulos, se consigue barrer toda la escena, pues lo que representa este plano son líneas de la escena, es decir éste atraviesa cada punto de la misma formando líneas paralelas entre sí, como en la figura (2.4).

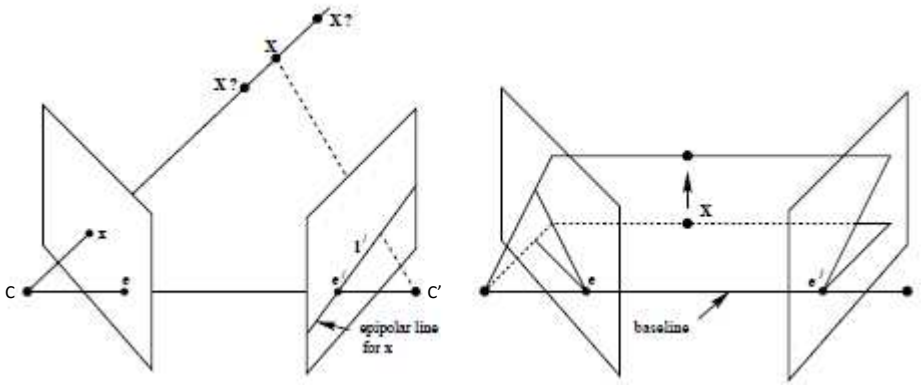

Figura 2.4. (Izquierda) Se observa el punto de interés X, en el plano imagen izquierdo el punto x generado por X sobre ese plano, en el plano imagen izquierdo está dibujada la línea epipolar l', en la cual debe encontrarse el punto correspondiente a x, el punto x'. (derecha) El plano epipolar con diferentes rotaciones para barrer toda la escena, el eje de rotación es el baseline.[11]

En la figura (2.4), los tres puntos x y x’ son la intersección entre el rayos que vuelven desde el punto X de la escena hasta los dos planos imagen respectivamente, como puede verse en la figura conocido el punto x, que está contenido en la línea epipolar de su plano imagen, el correspondiente punto x’ se debe encontrar en la línea epipolar del otro plano imagen, de manera que la búsqueda se reduce a una única línea (línea epipolar) en lugar de a toda la imagen.

Aún así sigue siendo necesario utilizar otros métodos para encontrar el punto correspondiente x’, para ello se hace uso de diferentes métodos de procesado, entre los cuáles se encuentran, el algoritmo de SIFT (Scale Invariant Feature-Transform), SURF(Speed Up Robust Features), ambos permiten detectar puntos relevantes en una imagen[27].

El principal inconveniente de este método es el de hallar la profundidad en imágenes con pocas texturas (homogéneas) o que contengan patrones repetitivos, ya que hace que la triangulación se convierta en un problema mal planteado, pues las correspondencias podrían hallarse erróneamente (muchas similitudes), una solución para el problema de superficies poco texturizadas es utilizar luz estructurada sobre la escena.

2.2. Luz estructurada

Otro de los principales métodos para la obtención de imágenes 3D es el uso de la "luz estructurada" , consiste en una iluminación activa de la escena mediante la proyección de un patrón que varía espacialmente en las coordenadas x e y . Este método también es llamado triangulación activa, por el hecho de utilizar una fuente de luz para iluminar la escena en lugar de adquirirlas con la luz existente en la misma.

Se puede considerar como una modificación del método estereoscópico explicado anteriormente, sustituyendo una de las cámaras por una fuente de luz encargada de proyectar la imagen (patrón), siendo tarea de la segunda cámara capturar la escena sobre la que está siendo proyectado dicho patrón. La proyección de un patrón que a priori es conocido simplifica el problema de correspondencia que existía en el caso de estereovisión con aquellas regiones con pocas texturas.

El principio de la captura de imagen 3D mediante este método consiste en extraer la información de la distorsión que presenta el patrón una vez proyectado sobre la escena , de manera que se obtienen dos imágenes , la capturada que contiene el patrón distorsionado por la geometría de la superficie de interés y el patrón proyectado originalmente.

Basándose en el mismo principio de medida que el método anterior, la triangulación geométrica (apartado 2.1), la distancia al objeto se determina con el conocimiento previo de los parámetros del sistema, como las posiciones de la fuente de luz y de la cámara que captura la imagen, de manera que el problema de correspondencia se reduce a identificar el patrón conocido que está siendo proyectado en la escena. En el caso más simple este patrón podría tratarse de una única línea de luz que debe desplazarse a lo largo de la escena, pero sólo es capaz de capturar una única línea por unidad de tiempo, a los dispositivos que utilizan este tipo de método se los denomina escáneres 3D pues, como se ha mencionado anteriormente la línea de luz debe recorrer la escena por completo al menos una vez para obtener su geometría, lo que implica que el sistema deba realizar al menos un barrido de la escena, haciendo que no sea posible su uso para objetos en movimiento.

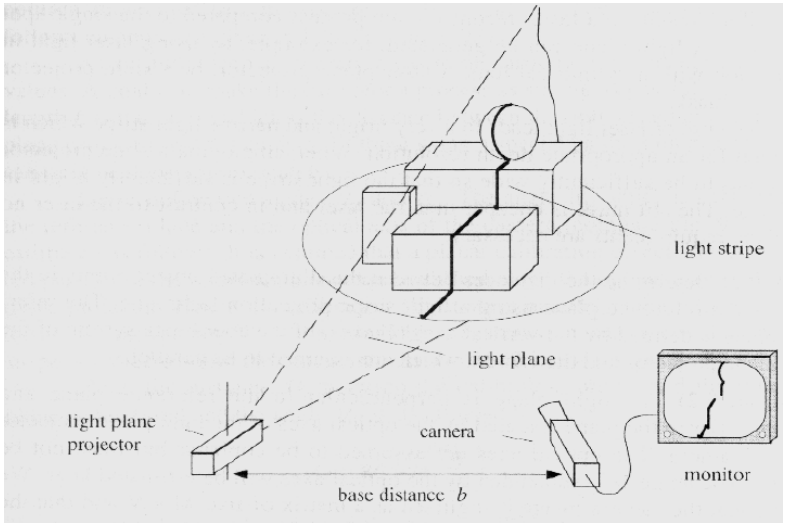

Figura 2.5. Adquisición de superficie mediante escaneado de escena. Puede verse cómo el proyector genera un único plano de luz que atraviesa la escena y debe realizar un barrido sobre la misma para capturarla por completo.

Con la proyección de una única linea de luz la correspondencia entre dos puntos dados en ambas imágenes era más fácilmente alcanzable, sin embargo para solventar los inconvenientes mecánicos que conlleva el anterior método, se opta por la proyección de un patrón con más elementos, teniendo más de una única línea (en el caso más simple) que proyectar. En función de las condiciones de la escena a capturar, existen diferentes tipos de patrones que pueden ser proyectados, tanto espacial como temporalmente.

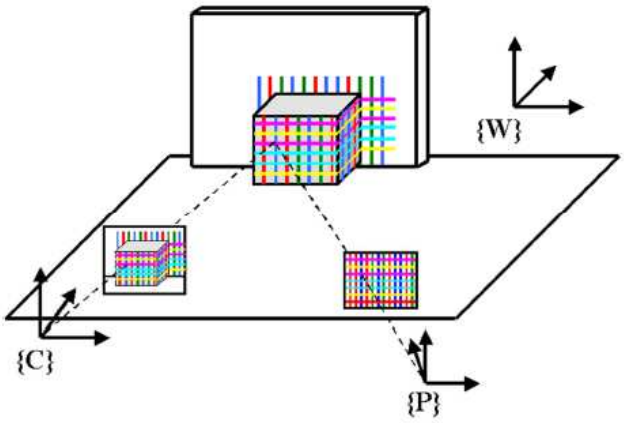

Figura 2.6. Técnica mediante luz estructurada, en la esquina inferior derecha se observa el patrón que está siendo proyectado, a la derecha se observa la imagen capturada por la cámara [5]

Estructura de los patrones

En función de las necesidades de la aplicación, se tienen distintos tipos de codificaciones (forma del patrón) y número de los mismos, es decir se puede utilizar más de un patrón para obtener la geometría de una superficie. Se pueden clasificar en:

-Multiplexados en tiempo: cada punto es codificado mediante una secuencia de intensidades. Codificación en tiempo.
-Codificación espacial: Cada punto es codificado por intensidades circundantes.
-Codificación directa: Cada punto es codificado por una única intensidad.

Multiplexación en tiempo

Consiste en una serie de patrones que son proyectados, de manera que cada punto de la imagen se identifica por una secuencia de intensidades recibidas.
La estructura más común de este tipo de patrones, consiste en una secuencia de líneas que se incrementan en número conforme avanza el tiempo.

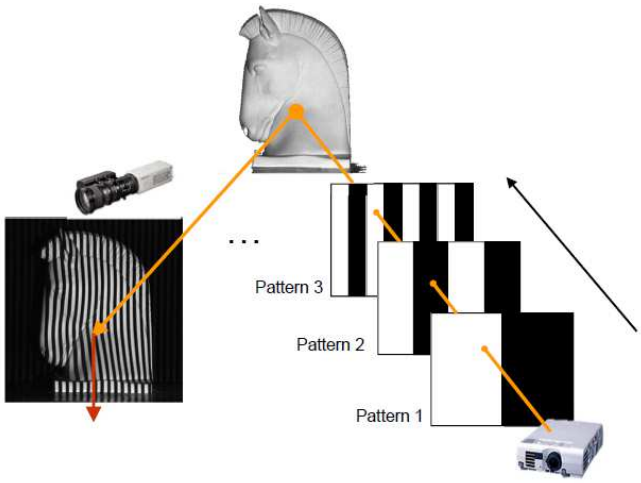

Figura 2.7. Técnica de multiplexación en tiempo, puede verse la proyección de varios patrones a lo largo del tiempo, que van creciendo en número de líneas [4]

Códigos binarios

Los patrones proyectados consisten en una serie de líneas en blanco y negro, el número de patrones determina el número de líneas que serán codificadas, de manera que n patrones deberán ser proyectados si se desea codificar 2^n líneas.

Códigos N-arios

Se trata de patrones en escala de grises, que reduce el número de patrones que serán necesario proyectar, incrementando los niveles de intensidades proyectadas (anteriormente sólo había 2), es decir incrementando la base del código.
El número de patrones, el número de niveles de gris y el número de líneas que contiene el patrón, están fuertemente relacionadas, fijando dos de estos parámetros se obtiene el resto.

A modo de ejemplo, si el objetivo fuera codificar un patrón de 64 líneas, con códigos binarios sería necesario 6 patrones ($2^6 = 64$).
Usando un código 4-ario, serían necesarios 3 patrones ($4^3 = 64$).

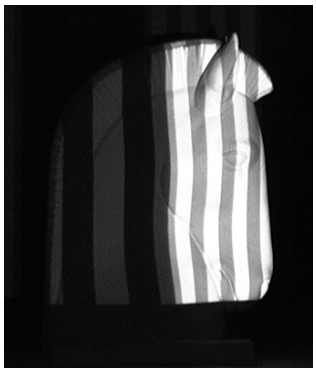

Figura 2.8. Códigos N-arios. Se observa que el patrón tiene diferentes tonos en escala de grises.[4]

Codificación espacial

La codificación espacial codifica, junto con los puntos de interés sus adyacentes, es decir se obtiene información de la región circundante al punto de interés, que se determina mediante una ventana. Esta codificación se realiza mediante un único patrón en lugar de multiplexar varios a lo largo del tiempo.
El objetivo de estas técnicas es obtener un sistema de medida mediante una única proyección, de manera que también se puedan medir superficies de objetos en movimiento.

Codificación directa

Este tipo de codificación también requiere únicamente la proyección de un patrón. Cada pixel es identificado por su propia intensidad o color, al ser necesaria una única proyección el espectro de intensidades o colores a utilizar para la codificación es muy amplio.
Entre las técnicas que se pueden encontrar de este tipo están:

Nivel de gris
Cada punto del patrón es identificado por su propio nivel de intensidad

Color
Cada punto se identifica por su color

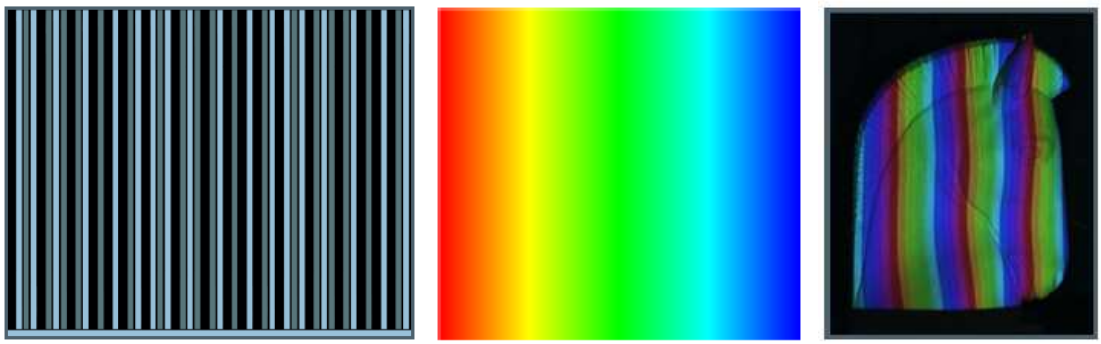

Figura 2.9. Nivel de gris (izquierda), patrón de color (color) , patrón de color proyectado sobre escena (derecha)[4]

Como se ha mencionado al principio de este apartado, la elección de una técnica depende del tipo de aplicación y de los requerimientos de la misma, a continuación se muestra una comparativa de las técnicas mencionadas anteriormente.

	Ventajas	Inconvenientes
Codificación temporal	-Alta resolución -Precisión de um -Robusto frente a objeto coloridos pues se usan patrones binarios	-Solo para objetos estáticos -Elevado tiempo de cómputo por el elevado número de patrones
Codificación espacial	-Apto para objetos en movimiento -Condensa la información en un único patrón	-Discontinuidades en la superficie del objeto, puede provocar decodificaciones erróneas (mala estimación de la profundidad)
Codificación directa	-Único patrón -Mayor resolución	-Se puede ver muy afectado por las propiedades reflectivas del objeto -Baja precisión (del orden 1mm)

Tabla 1. Comparativa de tecnologías de luz estructura.

A continuación se muestra una comparativa visual del funcionamiento de los tres tipos de técnicas vistas anteriormente.

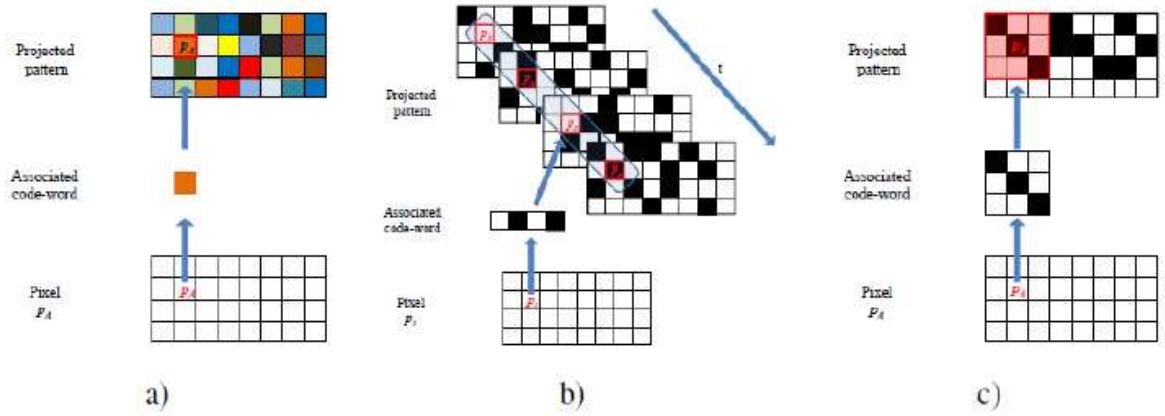

Figura 2.10. Comparativa técnicas de luz estructurada , a) Codificación directa, b) codificación temporal, c) Codificación espacial. Puede verse cómo sería la palabra código asociada a cada pixel [4]

La imagen a) corresponde a codificación directa, es decir el pixel del patrón proyectado (naranja), es el que debe buscarse el otra imagen, para proceder con la triangulación, ocurre lo mismo para b), pero en este caso la búsqueda del pixel correspondiente debe realizarse para varios patrones, que varían a lo largo del tiempo, por último en el caso c), no es únicamente el pixel de interés el que debe ser buscado en la imagen, sino también sus circundantes, a partir de ahí se puede proceder con la triangulación.

2.3. Tiempo de Vuelo (TOF)

Otra método muy extendido para la obtención de imágenes 3D, es el uso de las cámaras de tiempo de vuelo, que pueden ser usadas para estimar la estructura 3D directamente, es decir sin la ayuda de algoritmos de visión artificial, no existe problema de correspondencia como ocurría con los métodos anteriores.

Un sensor TOF (Time of Fligth) consiste en una fuente (proyector) de luz modulada, como un laser o LED, un sensor, que consiste en un array de pixeles, siendo cada uno de ellos capaz de detectar la fase de la luz entrante y un sistema óptico capaz de concentrar la luz en el sensor.

La medida de la distancia se consigue con la medida de la fase de la envolvente, tanto de la luz transmitida como la recibida en cada pixel del array.
Aunque en la realidad se utilizan ondas cuadradas para la modulación de la fuente, para exponer el modo de trabajo de ésta tecnología de manera sencilla, asumiremos que se utilizan ondas sinusoidales.

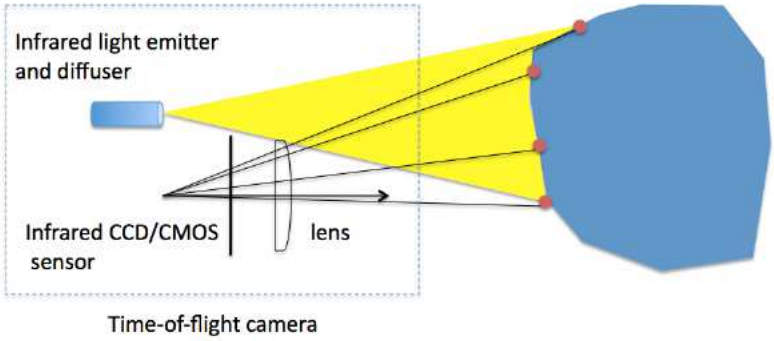

Figura 2.11. Sistema de vuelo. Que está formado básicamente por un emisor y difusor de luz, además el sensor infrarrojo con una lente para focalizar la luz. [7]

Existen dos tipos de técnicas para la medida del tiempo de vuelo, con sensor TOF de luz pulsada o de modulación continua.

Principio de funcionamiento: Con modulación continua

Sea s(t) la luz transmitida, donde f es la frecuencia de modulación, la luz reflejada que llega a un pixel del sensor con un desplazamiento ϕ, tendrá la forma :

$$s(t) = a.\cos(2\pi f t) \quad (3.1)$$
$$r(t) = A.\cos(2\pi f t + \phi) \quad (3.2)$$

Donde A es la amplitud de la señal reflejada, y ϕ es la fase de la señal recibida, de manera que la distancia al objeto se puede calcular como :

$$d = \frac{c\,\phi}{4\pi f} \quad (3.3)$$

Existe una ambigüedad en la medida debido a que de la distancia depende de la fase de señal recibida, que es periódica, con un periodo de valor c/2f. Por ejemplo para una frecuencia de 50MHz, la distancia máxima que se puede medir sin ambigüedad es de 3m.

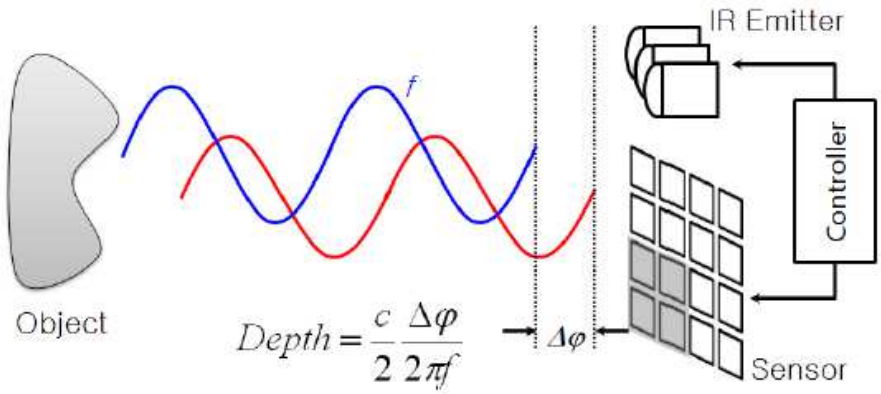

Figura 2.12. Modulación con luz modulada. [8]

A su vez el desplazamiento de fase se calcula atendiendo a la relación de cuatro cargas eléctricas diferentes, como se muestra en la figura (2.12). Estas cuatro señales tienen un desfase de 90º entre ellas.

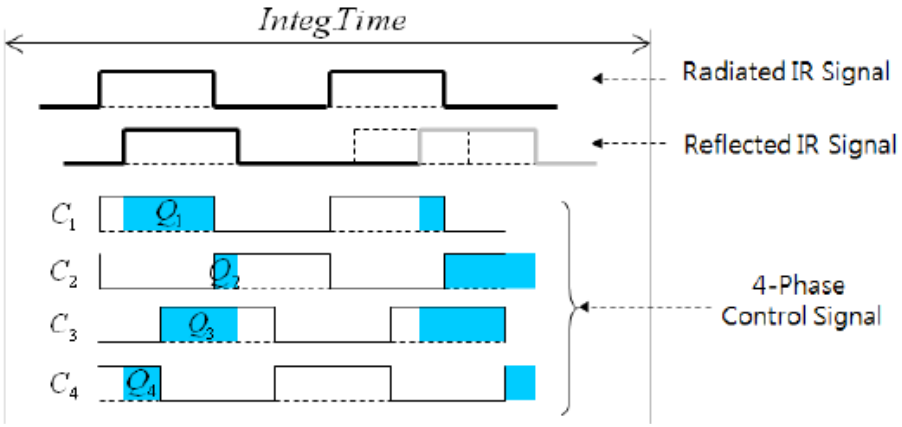

Figura 2.13. Señales sensor TOF. Señal enviada , señal recibida, C1 a C4 señales desfasadas 90º cuya carga obtenida por la señal recibida son respectivamente desde Q1 hasta Q4.[8]

De manera que la diferencia de fase se calcula, como una función de las cargas eléctricas que llegan al sensor (Q1, Q2, Q3, Q4).

Con el fin de mitigar los efectos de ruido en las medidas, se realiza un promediado de las medidas de distancia sobre varios periodos de señal, en la práctica el número de periodos comúnmente utilizados se sitúa entre 1 y 100 ms, por ejemplo para el caso de una frecuencia de modulación de fm = 30MHz, cuyo periodo es de $33{,}3\text{x}10^{-9}$ seg. , el número de periodos de modulación utilizado para realizar el promediado se encuentra entre $3\text{x}10^{4}$ y $3\text{x}10^{6}$ [28], a la longitud de este intervalo se le denomina tiempo de integración.
Con elevados valores del tiempo de integración, se consiguen medidas más precisas, aunque no se pueda aplicar para objetos en movimiento, pues requieren de periodos de integración reducidos.

Hasta este punto según el método de modulación de onda continua se puede concluir que:

-La cámara TOF trabaja a una frecuencia fija, lo que hace posible que múltiples cámaras trabajen simultáneamente, usando frecuencias diferentes.

-Con el fin de incrementar la relación señal a ruido, por tanto la precisión se necesitan unos periodos de integración relativamente largos, lo que introduce un suavizado en la imagen para objetos en movimiento.

<u>Principio de funcionamiento : fuente pulsada</u>

El laser genera un pulso de luz de unos pocos nanosegundos, la distancia es calculada directamente del retardo entre el pulso emitido y su reflexión.
El ancho del pulso de luz es de unos pocos nanosegundos, lo que produce un pulso de elevada potencia óptica, permitiendo al dispositivo trabajar en exteriores, en condiciones adversas [6]. Están basados en diodos de avalancha de fotones (SPAD), por su habilidad de detectar fotones individualmente con una alta resolución en tiempo de llegada, aproximadamente de 10 picosegundos.

Debido a que no hay modulación de envolvente, no existe problema de ambigüedad de fase, pues se mide directamente el tiempo de llegada tampoco de tiempo de integración.

Es la tecnología elegida para numerosas aplicaciones en exteriores bajo condiciones adversas como topografía (estática y móvil), conducción autónoma, misiones planetarias como la creación del robot humanoide R2 (NASA)[1]

[1] http://www.mvtec.com/news-press/article/detail/vision-guided-robot-heads-for-space-nasa/

3.-Cámaras de profundidad

Los dispositivos utilizados para la captura de imágenes 3D se denominan cámaras de profundidad.
Las principales tecnologías que utilizan estos dispositivos, son visión estereoscópica, luz estructurada y tiempo de vuelo, explicadas en párrafos anteriores.
Entre los modelos que existen en el mercado, según el tipo de tecnología empleado se encuentran los siguientes modelos:

Visión estereoscópica : - Panasonic AG-3DA1 [1]

Cámaras de luz estructurada :

-Kinect
-Asus Xtion Pro live
-Ensenso (http://fr.ids-imaging.com/ensenso.html)

Tiempo de vuelo:

- Luz modulada SR4000 (Swiss Ranger) (http://fr.ids-imaging.com/ensenso.html)
-Luz pulsadaTiger Eye (Advanced Scientific Concepts, Inc.)
(http://www.advancedscientificconcepts.com/index.html)

[1] http://www.panoramaaudiovisual.com/2010/02/17/panasonic-da-a-conocer-las-especificaciones-para-su-camara-ag-3da1/

Figura 3.1 Arriba (izquierda) Kinect , arriba (derecha) Ensenso, centro (izquierda) SR4000 Swiss Ranger, centro (derecha)Tiger Eye[7], abajo (izquierda y derecha) Panasonic AG-3DA1 [www.avtiimi.fi]

3.1-Elementos principales

Los elementos principales de una cámara de profundidad son:

-Cámara RGB de baja resolución
-Cámara infrarroja
-Proyector infrarrojo

Figura 3.2. Cámara Asus Xtion Pro live, con sus componentes principales señalados [12]

3.2.Principio de funcionamiento

Para conocer la distancia de un objeto de la escena hasta la cámara, es necesario además tener un proyector, que en el caso del modelo de trata de un proyector infrarrojo, por lo que la cámara de que captura la escena debe serlo también. El funcionamiento es el siguiente.

1. Se enciende el proyector de luz infrarroja infrarroja , que proyecta un patrón irregular de puntos, con una longitud de onda de entre 700 nm y 1 mm, que pertenece al espectro de luz infrarroja el cual no es visible para el ojo humano. Es posible visualizar este patrón con una cámara de visión nocturna.
2. Con el patrón proyectado en la escena, la cámara infrarroja será la encargada de capturar la luz infrarroja devuelta a la cámara, pues a diferencia de las cámaras RGB, las cámaras infrarrojas contienen sensores CMOS que pueden detectar la luz infrarroja que “rebota” de todos los objetos que se encuentran en la escena.

Figura 3.3. Patrón de puntos proyectado por cámara infrarroja , observado mediante una cámara de visión nocturna, es el tipo de patrón que proyecta la cámara utilizada para el proyecto[12].

3.En este punto se calcula la profundidad para pixel en la escena, esto se realiza tomando sendas imágenes, la captada por el sensor y el patrón de puntos proyectado inicialmente, hallando los puntos correspondientes entre ambas imágenes, para posteriormente utilizar triangulación y así calcular la distancia final al objeto.

3.3. Imagen de distancias o mapa de profundidad

El aspecto de una imagen de distancias es el mostrado en la figura (3.4) , donde se puede observar una imagen en escala de grises, que presenta valores pequeños (pixeles oscuros), en las zonas donde se encuentran objetos más cercanos a la cámara y valores más altos para aquellos objetos más alejados(en la caso del ejemplo : pared), las zonas negras de la imagen corresponden a fallos a la hora de la captura de la imagen. En la imagen se observa, zonas negras correspondientes a las estantería y reposabrazos de la silla, que como se observa en la imagen izquierda, son oscuras, por tanto no reflejan la luz, las sombras cerca de la persona, se deben a problemas de oclusión, pues en esas zonas, el sujeto está impidiendo que el patrón se proyecte sobre ellas, por lo que la cámara no detecta patrón esas zonas. Además de los problemas que se muestran en las capturas de las imágenes, se podrían dar errores en la medida en escena que contengan objetos especulares, o que el sujeto se encontrara demasiado alejado, o por el contrario demasiado cerca.

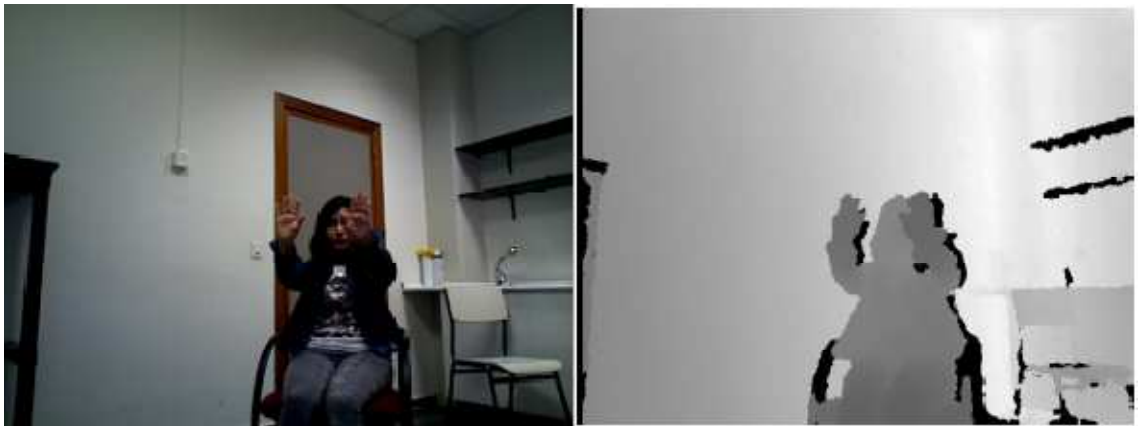

Figura 3.4. Imagen original (izquierda), imagen distancias obtenida con cámara de profundidad (derecha)

4. Técnicas de tratamiento de imagen

4.1.Morfología

La morfología es un procesamiento señal basada en máximos y mínimos. La erosión y la dilatación son las técnicas más utilizadas de este tipo de procesamiento.

Elemento estructural

El elemento estructural es en morfología matemática lo que una máscara (o núcleo) de convolución es en los filtros lineales. Tiene un centro (o anchor point).

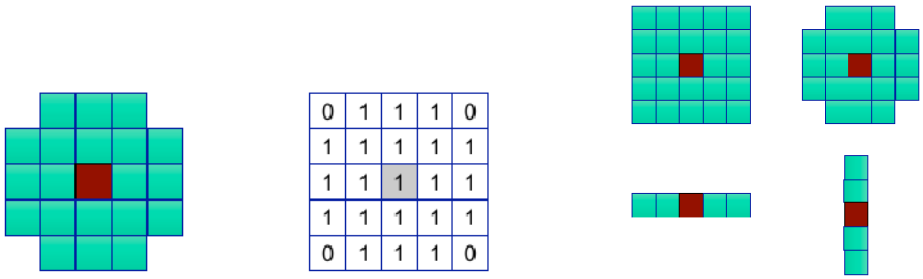

Figura 4.1. Elemento estructurante, representa la matriz (derecha), el anchor point está en el centro, ésta máscara se utiliza para recorrer toda la imagen, colocando se dicho centro en cada uno de los píxeles de la imagen [10].

Erosión

La erosión consiste en la expansión de las zonas oscuras de una imagen. Para el proceso se va desplazando el elemento estructurante por todos los píxeles de la imagen, colocando el centro del EE sobre cada uno de estos. En la imagen resultado el valor de cada uno de los píxeles será el mínimo de los píxeles, que en la imagen original estén bajo la máscara.

Figura 4.2. Detalle erosión. Se observa que al desplazar la máscara por todos los píxeles de la imagen (izquierda) el resultado (derecha) obtenido finalmente es una ampliación de las zonas oscuras.

Dilatación

Es un proceso similar al de erosión, pero en este caso lo que se produce es una extensión de las zonas claras, este caso tomando los máximos bajo la máscara.

Figura 4.3. Detalle dilatación. Se observa que al desplazar la máscara por todos los píxeles de la imagen (izquierda) el resultado (derecha) obtenido finalmente es una ampliación de las zonas claras (o reducción de las zonas oscuras)

4.2- Segmentación

Los diferentes objetos que pueden aparecer en una imagen, están formados por un conjunto de píxeles, para identificarlos es necesaria la agrupación de píxeles con características similares, pues cada uno ellos presenta características como:

-Color
-Movimiento
-Textura

Existen métodos como el de la segmentación de componentes conexas, que consiste en agrupar píxeles conexos.
La segmentación por zonas planas, consiste en agrupar píxeles con características similares, como las enumeradas anteriormente, que además está formada por píxeles conexos. Esta es la que se aplica en el proyecto.

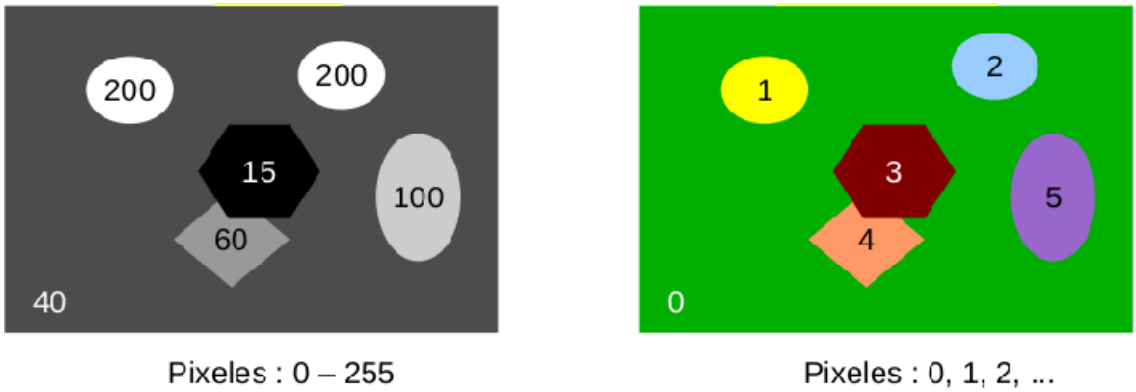

Figura 4.4. (Izquierda) Imagen original en escala de grises, donde los números representan el nivel de gris de cada objeto. (Derecha) Imagen tras la segmentación, se observa que se ha sustituido el nivel de gris, el valor del nivel de gris por una etiqueta, que es la que identificará a cada uno de los objetos de la imagen, también se observa que el fondo ha tomado el valor de 0, pues este es el valor que se suele asignar en estos caso. Los colores asignados a los objetos que se observan en la imagen derecha tienen un fin únicamente visual, ya que se trabajar con los valores de los pixeles de las etiquetas.[10]

5.Objetivo: Reconocimiento de gestos

Como se ha introducido anteriormente el objetivo del presente proyecto es el reconocimiento de los gestos, realizados por un sujeto colocado delante de una cámara de profundidad. Tras haber expuesto algunas de las técnicas de captura de imagen 3D, los elementos principales de una cámara de profundidad y su principio de funcionamiento, además de algunas técnicas de procesado de imagen, precisamente las más importantes empleadas en este proyecto, en los siguientes apartados se expondrán las diferentes etapas que se han empleado para la consecución de dicho objetivo.

Previamente es necesario conocer el tipo de gestos que se van detectar, por lo que se va a analizar la variación de sus coordenadas en pixeles a lo largo del tiempo.

Naturaleza de los gestos

El tipo de gestos que se van a detectar, contienen movimiento en una única dirección según el gesto, en concreto se trata de gestos en los que las manos únicamente realizan movimientos horizontales o verticales, según corresponda.
Los movimientos que se van a analizar constan de 3 etapas. Las etapas se denominan inicial, movimiento, y final, en las etapas inicial y final el sujeto permanece casi estático pues se trata del comienzo y la finalización del movimiento.

A continuación se van a mostrar los gestos que se han capturado en las secuencias, y las etapas de cada movimiento.

Por la similaridad de los movimientos entre sí, se analizarán dos movimientos horizontales.

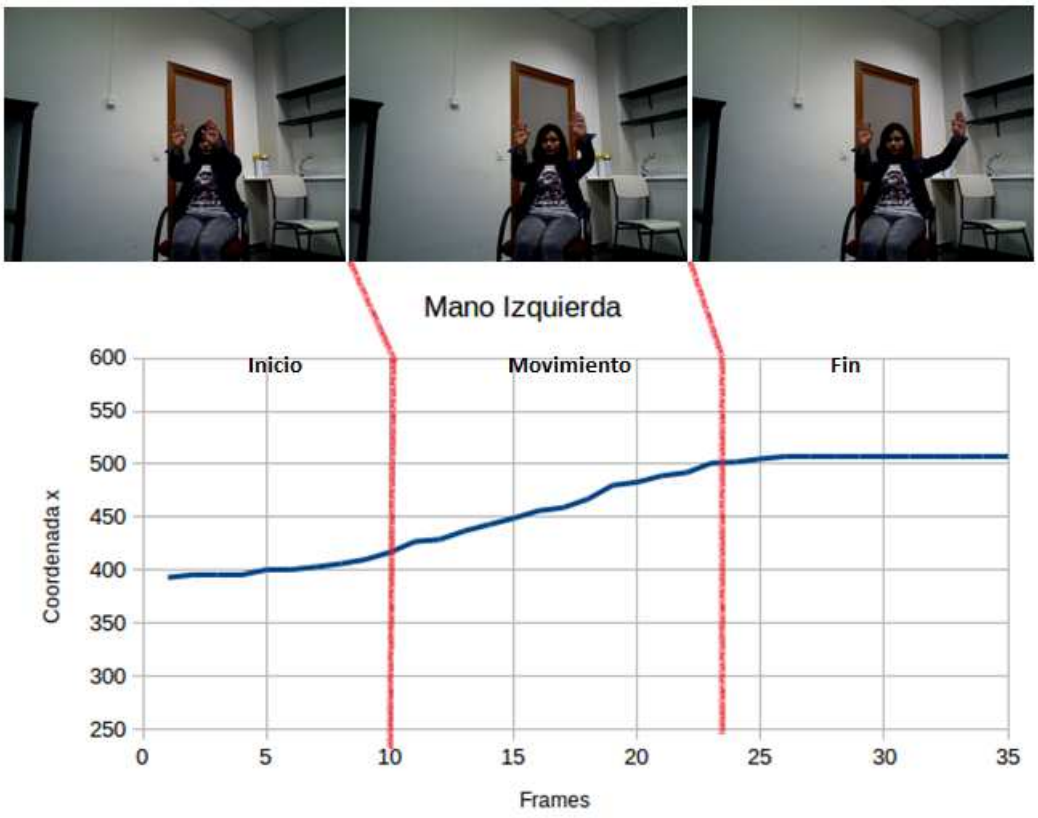

Figura 5.1.a) Movimiento de mano izquierda (arriba) y gráfica de variación de coordenadas x

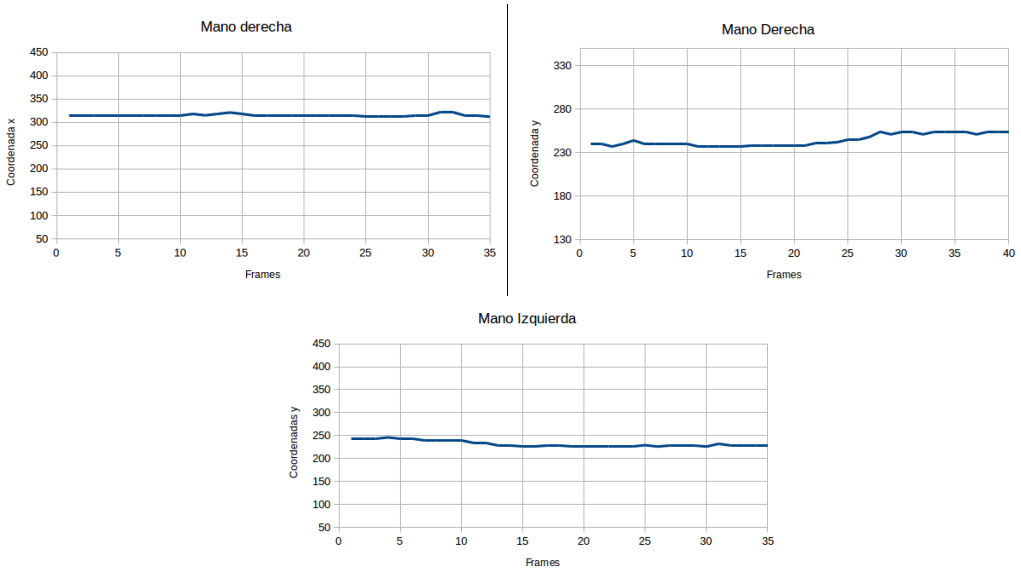

Figura 5.1.b) Gráficas de coordenadas x e y mano derecha (arriba) y coordenadas y de mano izquierda (abajo)

En la figura (5.1. a)), se puede observar como el movimiento se realiza variando únicamente las coordenadas x de la mano izquierda, también se diferencian las tres etapas del movimiento:

-El momento del inicio, la gráfica apenas varía, esto es debido a que las manos no están perfectamente estáticas, si así fuera esa zona tendría la forma de una línea recta horizontal.

-Movimiento, este es el espacio temporal, en el que la manos comienzan a moverse, su desplazamiento es hacia la izquierda, por lo que el valor de la coordenada x, aumenta.

-Final, momento en el cual el gesto ha llegado a su fin, tiene las mismas características, que el momento de inicio.

Se debe permanecer un tiempo en el último estado antes de iniciar otro gesto, de lo contrario ninguno de los dos sería detectado.

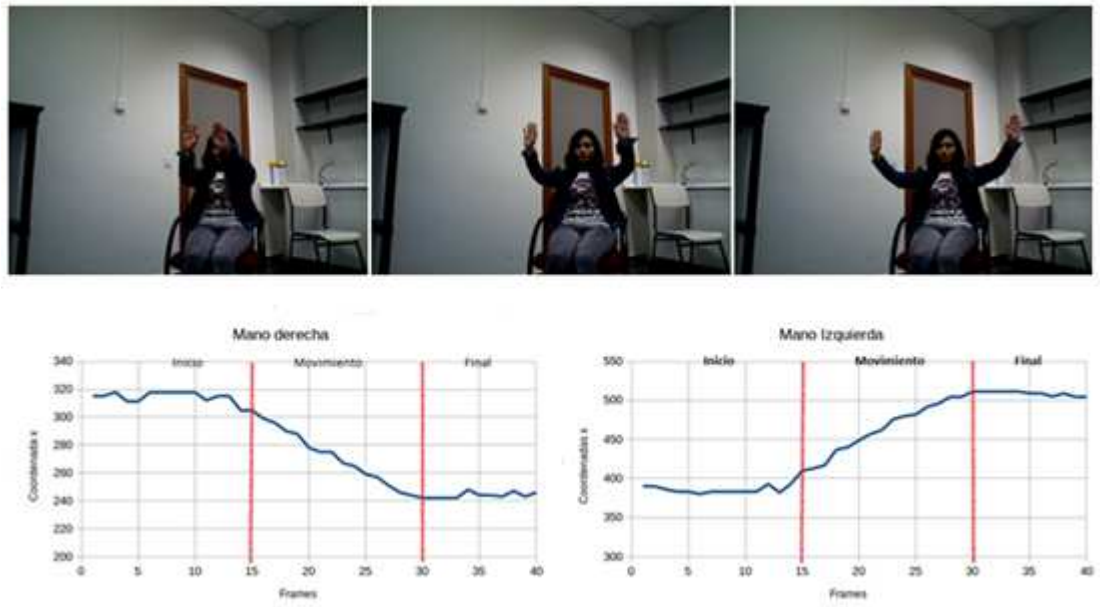

Figura 5.2.a) Movimiento de ambas manos en modo expansión (arriba) y gráficas de variación de coordenadas x.

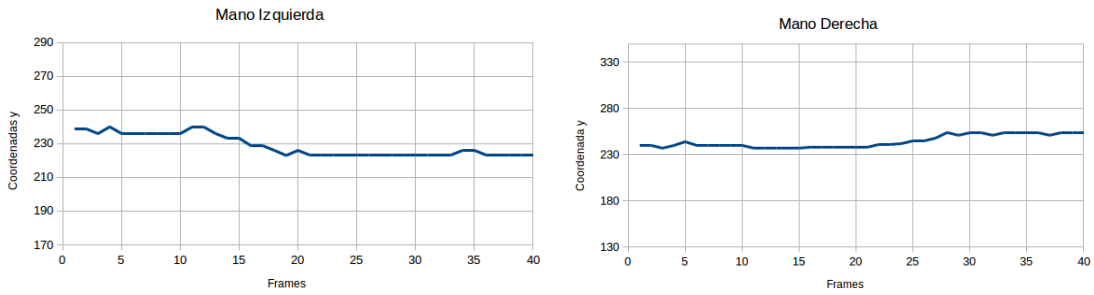

Figura 5.2.b) Gráficas coordenadas y, izquierda y derecha.

En el gesto de la imagen (5.2.a)), ambas manos realizan el movimiento, que también es en sentido horizontal, por lo que, como se observa en la gráficas de las coordenadas x de ambas manos, las fases del movimiento son las mismas para cada mano, como ocurría en el gesto anterior (5.1.a)), en este caso destacar que para el caso de la mano derecha la coordenada x disminuye, pues el movimiento es a la derecha (las coordenadas de los píxeles son tales, que el pixel de la esquina superior izquierda tiene las coordenadas (0,0)).

Etapas de diseño

Para la elaboración del trabajo, se han planteado tres sub-objetivos, cada uno de los cuales comprende una etapa de diseño con sus correspondientes bloques, la función de estas etapas se explicarán en los siguientes párrafos.

5.1-Detección de objetos (manos)

Se trata de la primera etapa de desarrollo del programa, pues se trabaja directamente con la imagen de distancias captada por la cámara de profundidad. En ella cabe distinguir dos bloques, el de "detección de mínimos" y "clasificador de mínimos".

5.1.1-Detección de mínimos

Este bloque se ha realizado mediante la creación de una función, que emplea el método de erosión, técnica explicada anteriormente. El objetivo a partir de aquí, es detectar que objetos de la escena que podrían ser manos.

Figura 5.3. Imagen RGB (izquierda), Imagen de distancias (derecha)

Tomando como ejemplo la imagen 5.3, obtenida de una de las secuencias de prueba empleadas, se observa que el sujeto está sentado en una silla frente a la cámara con las manos levantadas y alrededor hay otros objetos como estanterías y sillas. Para comenzar con la detección del gesto, en primer lugar se debe realizar la detección de las manos, para ello se deben encontrar los mínimos de la imagen, entre los cuales deben encontrarse dos, los correspondientes a las manos.

Preprocesado de imagen de distancias

Como se ha mencionado anteriormente debido a fallos en la medida, aparecen zonas negras en la imagen, cuyos pixeles tienen un valor de cero. La zona de interés, es decir aquella donde debe situarse el sujeto para la correcta detección de las manos, está alrededor de 1.5 m, además se observa que detrás del sujeto hay una pared, la cual está situada a unos 3 metros, valor elegido como máximo en la imagen. Por ello se va modificar la imagen de distancias antes de empezar el proceso de detección de mínimos.

-En primer lugar se recorta el valor de los píxeles de la imagen al valor máximo elegido, de manera que resultaría una imagen como la de la figura (5.4 (centro)), pues los objetos próximos a la pared no serán de interés ya que el sujeto no va a situarse en esa zona.

- A continuación se sustituyen los píxeles negros por el valor máximo. La imagen resultante es (5.4 (derecha)).

Figura 5.4. Imagen de distancias original (izquierda), imagen de distancias acotada en máximos (centro), imagen de distancias corregida (derecha)

Una vez finalizado el preprocesado de la imagen, ya se puede comenzar con la detección de mínimos.

Detección de mínimos

-El primer paso es realizar la erosión de la imagen, como se ha explicado anteriormente, esto se hace para expandir las zonas oscuras de una imagen. El elemento estructurante elegido finalmente, tras realizar pruebas sobre diferentes fotogramas ha sido un elemento rectangular con una componente vertical dos veces mayor que la horizontal. De manera que tiene unas dimensiones pensadas para que el mínimo correspondiente a una mano no interfiera en la otra, es decir que se puedan detectar los dos mínimos.

Figura 5.5. (Izquierda) Imagen distancias con las manos señaladas mediante círculos rojos, (derecha) Resultado de la erosión de la imagen de distancias, se marcan los puntos correspondientes al lugar donde están las manos, se observa que el esparcimiento de las zona más oscuras que corresponden a las manos.

- Una vez realizada la erosión, se procede comparar sendas imágenes, la original y la erosionada. Tras realizar la erosión los valores mínimos de la imagen se han expandido a zonas circundantes, en la imagen resultado sólo estarán los píxeles que coincidan en ambas imágenes, que corresponderán con los mínimos.

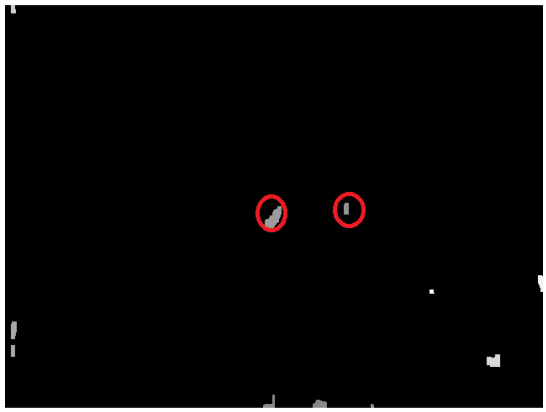

Figura 5.6. Mínimos de la imagen, los puntos señalados en rojo corresponden a las manos

4. Como se puede observar en la imagen resultado (5.6), hay varias zonas consideradas mínimos pero también se observar que la mayoría se encuentran en los laterales de la imagen que no serán considerados zonas donde se podría situar el sujeto, pues ha de situarse preferentemente alrededor de la zona central. Por tanto se procede a "limpiar" la imagen para quedarnos únicamente con la zona central, a partir del resultado procederemos con el siguiente bloque.

Figura 5.7. Mínimos de la imagen, tras la eliminación de mínimos que no podrían ser considerados manos.

5.1.2-Clasificador de mínimos

Partiendo del resultado obtenido en el bloque de "Detección de mínimos", este bloque será el encargado de decidir sí los mínimos existentes en una imagen corresponden o no a una pareja de manos, según una serie de criterios que deberán cumplir.
Un paso previo antes de empezar a clasificar dichos mínimos, es segmentar la imagen, para lo cual aplicaremos el método de segmentación por zonas planas que se ha expuesto anteriormente, de manera que tendremos los mínimos de la imagen etiquetados.
Como se puede observar en la imagen (5.7), los mínimos son puntos en la imagen que corresponden a las manos, pero para poder clasificarlos, necesitamos obtener su forma real, o dicho de otra manera debemos extraer de la imagen original todo el contorno del objeto.

Para realizar esta labor se ha creado una función que rastrea píxeles semejantes a uno dado. Conocidas las coordenadas del pixel de interés, que son los mínimos, se almacenan las coordenadas de aquellos píxeles que se encuentran en las proximidades del punto, considerando que son pixeles que pertenecen a las manos. La decisión de si un pixel pertenece o no a la mano, se toma en base, a que la diferencia entre el mínimo correspondiente a ésta y la del pixel evaluado no supere cierto umbral. En la siguiente imagen (5.8)se puede ver el resultado tras aplicar la función a la imagen de mínimos.

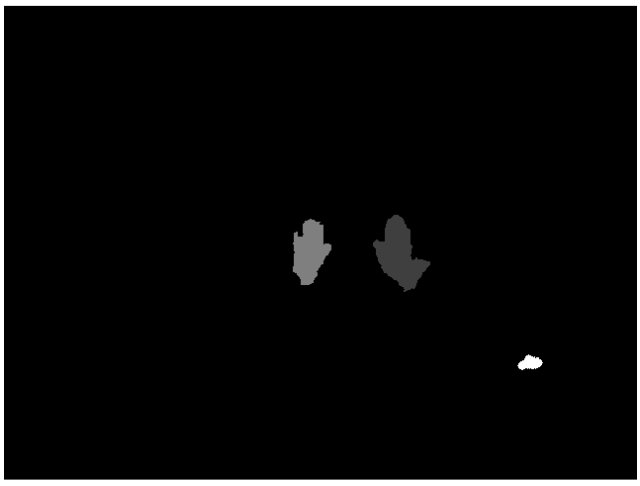

Figura 5.8. Extracción del contorno de las manos de la imagen de distancias, puede observarse que hay también otros objetos, estos serán descartados en pasos posteriores.

Los criterios según los cuales se va a decidir si en la escena hay un par de manos frente a la cámara, serán:
-Las dimensiones, pues como es evidente, no puede haber manos excesivamente grandes ni excesivamente pequeñas.
-La separación entre ambas manos, que cambia como es evidente en función del sujeto a prueba, pero tampoco debe ser una distancia demasiado elevada.

Para asegurarse de que los objetos cumplen las condiciones referentes a las dimensiones, y separación entre manos, se ha creado una función para cada una de ellas, de manera que los objetos deben cumplir ambas condiciones para ser consideradas manos, cumpliendo en primer lugar la condición referente a dimensiones.

Una vez se han detectado manos en la escena, se asume que delante de la cámara hay un sujeto, que va a realizar un gesto, por tanto una vez finalizado este bloque, se pasa bloque de detección de movimiento, de lo contrario se siguen analizando las imágenes de distancias a la espera de la detección de manos.
El resultado final de este proyecto, es puesto a prueba mediante unas secuencias, cada fotograma de la secuencia debe pasar por este bloque, obteniendo finalmente para cada mano detectada, sus coordenadas 3D.

5.2. Detección de movimiento

Una vez realizada la detección de las manos, mediante el bloque anterior, para detectar si el sujeto que está delante de la cámara está realizando un gesto o no, es necesario comprobar si las manos están realizando algún tipo de movimiento, que pueda ser reconocido como un gesto, que el programa sea capaz de detectar. Para ello se ha implementado el siguiente bloque, que a su vez, está compuesto de dos bloques.

Frame a frame se obtienen las coordenadas de puntos correspondientes a ambas manos y se almacenan en arrays diferentes, uno para cada mano.

5.2.1. Bloques de movimiento

Los gestos tendrán lugar durante una serie de frames consecutivos, generándose puntos 3D, en cada uno de ellos. Para que se perciba una variación considerable de un frame al siguiente, el sujeto debería realizar gestos muy rápidos, además al no permanecer perfectamente estático mientras no está realizando ningún movimiento, los puntos oscilan alrededor de una posición. Por este motivo para analizar los gestos, se van a utilizar una serie de bloques, que llamaremos bloques de movimiento, cada gesto estará compuesto por una serie de estos bloques.

Cada uno de estos bloques se crea mediante una serie de puntos 3D, en concreto 7, en apartados posteriores se explicará las razones de esta elección. Tomando el primer y el último punto de estos, se calcula su distancia euclídea y se compara con cierto umbral, si esta distancia lo supera, se considera que es un bloque de movimiento, de lo contrario será un bloque sin movimiento.

Estos bloques también deben ser almacenados, para posteriormente ser analizados, por el siguiente bloque de procesamiento.

Con los bloques de movimiento obtenidos, cuyo valor es de tipo lógico pues es ‘1’ si ha habido movimiento y ‘0’si no lo ha habido, obteniéndose una secuencia de ‘1’s’ y ‘0’s’. Para realizar el

análisis de estas secuencias se van utilizar las fases del gesto explicadas en párrafos anteriores. Para reconocer estas fases del gesto se hace uso del bloque llamado 'Analizador de patrones'.

5.2.2.Analizador de patrones

Este bloque es el encargado de analizar las secuencias binarias para detectar gestos, la manera de hacerlo será comparar estas secuencias con unos patrones, que serán generados dinámicamente en función de la entrada al bloque.
Debido a que un gesto debe tener una duración mínima, que evidentemente puede variar en función de la rapidez del sujeto en realizar el gesto o del tipo de gesto en sí mismo, se ha fijado en este caso en 1 segundo, pues en el caso de estudio es el tiempo mínimo empleado para realizar un gesto(en el caso de la cámara que está siendo utilizada 25 fps).
En primera instancia serán necesarios un mínimo de cuatro bloques para poder determinar si ha habido gesto o no, pues corresponden a 28 fotogramas (múltiplo de 7), que es aproximadamente un segundo de secuencia y según el párrafo anterior, la duración mínima que se ha considerado que ha de tener un gesto.

Teniendo en cuenta lo mencionado anteriormente sobre las partes que conforman, un gesto de duración mínima, obtenemos el patrón con el cual se debe comparar la secuencia como sigue, asignando el valor de bloque que corresponde :

-Bloque 1 : Momento inicial (semi estático - movimiento): 0
-Bloque 2 : Instantes correspondientes al movimiento efectivo de la mano (movimiento - movimiento): 1
-Bloque 3 : Instantes correspondientes al movimiento efectivo de la mano (movimiento - movimiento): 1
-Bloque 4 : Momento final del gesto (movimiento - semi estático) : 0

De esta manera se ha obtenido el patrón de mínima longitud al que se deberán ajustar los bloques de entrada al analizador (Patrón: 0 1 1 0). Puesto que podría tratarse de un gesto de mayor duración, o que el sujeto tarde más tiempo en realizarlo, el patrón debe extender su longitud dinámicamente, es decir en función de la secuencia de entrada, pudiendo llegar a tener la siguientes longitudes y formas:
- 0 1 1 1 0
- 0 1 1 1 1 0

Como se puede observar se trata de patrones con la misma estructura que el de tamaño mínimo pero que contienen en la parte central del patrón un mayor número de bloques de valor 1. Estos serán los patrones utilizados para detectar gestos de mayor duración y por tanto la parte correspondiente al movimiento efectivo de la
mano es más extensa.

Una vez que ya se tienen los cuatro bloques que se necesitan mínimamente a la entrada del analizador de patrones, se procede a comparar bloque a bloque la secuencia de entrada, con el patrón de mínima longitud que hemos obtenido anteriormente, han de coincidir todos y cada uno de los bloques con los correspondientes bloques del patrón.

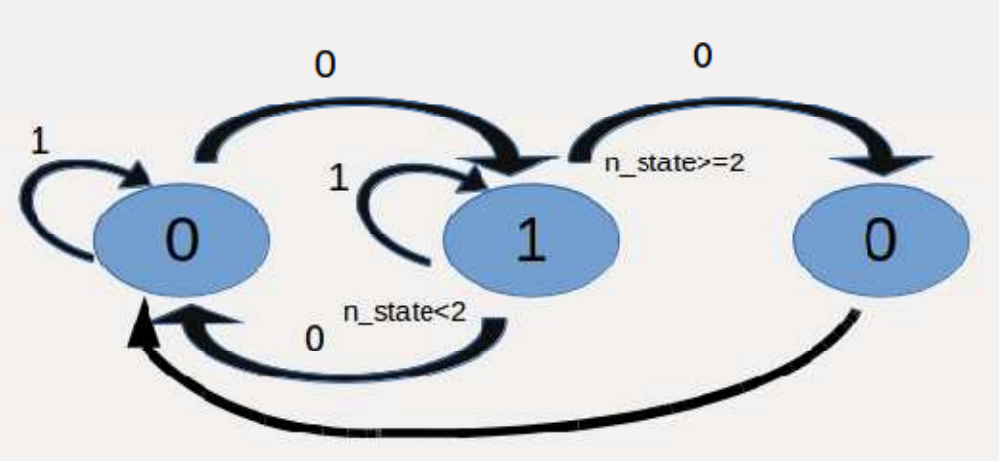

Figura 5.9. Diagrama de estados del analizador de patrones, la variable n-state contiene el número de transiciones que se ha permanecido en el estado.

La comparación del patrón con las secuencias se ha realizado mediante una maquina de estados, que está formada de tres estados:

-Estado inicial ('0') : Corresponde a la parte estática al inicio del gesto.
-Estado de movimiento ('1') : Corresponde a la parte dinámica del gesto.
-Estado final ('0') : Corresponde a la parte final del gesto y el último de los estados por los que debe pasar la secuencia.

Su funcionamiento es el siguiente :

-Estado inicial : Se parte de que el primer bloque de la secuencia debe ser igual a '0', que corresponde al primer bloque del patrón, si la comparación resulta verdadera, se avanza al estado de movimiento, de no ser así, el estado siguiente debe ser el inicial.

-Estado de movimiento: Si se llega a este estado, significa que el primer bloque de la secuencia que podría corresponder con gesto tiene el valor de '0', por tanto el siguiente debe tener un valor igual a '1', que corresponde con la parte de movimiento. Realizamos la comparación al igual que en el estado anterior. Como se ha mencionado en apartados anteriores la duración del gesto podría variar y ser por tanto mayor al patrón mínimo (nunca menor), lo que implica que la secuencia potencialmente detectada como gesto tendría longitud mayor, y como se han visto esto corresponde a un aumento en el numero de bloques correspondientes a movimiento para poder cumplir con el patrón. Esto implica que se deba permanecer en este estado durante varias transiciones, para el caso de mínima longitud del patrón se deberá permanecer en este estado durante dos transiciones, que será también el número mínimo de transiciones que se deberá permanecer en este estado. A modo de aclaración permanecer en el estado significa que el bloque que está siendo analizado cumpla la condición de permanencia en el estado, que es como se ha dicho antes que su valor sea igual a '1', de no ser así se debe volver al estado inicial. Llegados a este punto, prácticamente ya se puede considerar que se ha realizado un gesto, pues únicamente se debe esperar a que se deje de cumplir la condición de permanencia en el estado para pasar al siguiente, que corresponde a la parte final del gesto.

-Estado final : La llegada a este estado, significa que hemos cumplido los requisitos de duración mínima de un gesto y además éste, ya ha llegado a su fin.
En este estado no hay condición de permanencia, pues el siguiente estado será siempre el estado inicial, donde se repetirá el proceso desde el principio.

Una vez se ha detectado un gesto, se toma nota del bloque de inicio y final de la secuencia del gesto, que posteriormente serán utilizados para el reconocimiento del gesto realizado. Además, puesto que el programa se ha implementado por bloques, solo aquellos objetos que sean detectados como manos y hayan coincidido con alguno de los patrones, pasaran al siguiente y último bloque, por tanto el analizador devuelve un valor lógico a por objeto, que tomara un valor de '1' en caso de haber seguido algún patrón o '0' que en caso contrario, que significa que el objeto es una mano pero que no está realizando ningún movimiento. Como es de esperar debe haber al menos un objeto que tenga asociado un valor lógico de '1', o en caso negativo no se pasa al siguiente bloque.

5.3.Reconocimiento de gesto

Este es el último bloque del diseño del programa, pues es el encargado de que una vez se ha decidido que se está realizando un gesto, identificar de qué tipo de gesto se trata. Aunque a priori podría parecer complicado, pues en función del tipo gesto que se pretende detectar podrían ser necesarias operaciones algo más complejas, sin embargo debido a que el tema central del proyecto, es presentar la utilidad de las imágenes profundidad, el objetivo del programa se ha centrado en reconocer gestos sencillos, verticales y horizontales, pues como se expone en párrafos anteriores, se trabaja todo el tiempo partiendo de las coordenadas 3D de los objetos que se han detectado como manos, de manera que, conociendo la trayectoria de dichos objetos o dicho de otra manera, la trayectoria de los puntos correspondientes a estos, se puede deducir el tipo de gesto realizado.
Las coordenadas 3D almacenadas anteriormente, serán útiles ahora también para identificar el tipo de gesto realizado, pues en el analizador de patrones se ha tomado nota del punto de inicio y final del gesto.

El funcionamiento es el siguiente:

- En primer lugar se ha de comprobar, si el gesto se realiza moviendo una sola mano (la otra permanece en la posición inicial de las manos) o con las dos, para ello se comprueban los dos valores lógicos que el analizador de patrones devuelve. Por tanto a partir de aquí ya se ha hecho una primera selección de entre dos posibles grupos, gestos con las dos manos y gestos con una sola mano.

- Posteriormente se han de obtener las secuencias de puntos correspondientes al gesto que se está realizando, esto se lleva a cabo mediante la utilización de los pares valores de valores devueltos por el analizador de patrones, que indican el inicio y el final del movimiento que está realizando cada mano, estos valores son los índices que corresponden a los puntos dentro de los arrays.
Antes de pasar al siguiente paso, es necesario aclarar que una vez se ha hecho uso de la información contenida en la coordenada z, y como se ha expuesto anteriormente los gestos que se han planteado detectar tienen únicamente componente vertical u horizontal (x o y), estas serán las coordenadas que se utilizaran en el siguiente paso.

-Llegados a este punto únicamente queda determinar si el movimiento realizado es horizontal o vertical, tanto si el movimiento se ha realizado con una sola mano o con las dos el procedimiento que se ha seguido para determinar la dirección es el mismo, la generación de unas variables (deltas) horizontales y verticales, es decir la cantidad de movimiento que ha habido en un sentido o en otro, que se han obtenido calculando la diferencia entre las coordenadas x e y respectivamente, entre el inicio y el final de la secuencia de duración del gesto de los vectores de puntos correspondientes. Como se ha explicado al principio de la sección debido a la sencillez de los gestos ha detectar, mediante la comparación de ambas deltas, horizontal y vertical, podremos determinar cuál es mayor y por tanto también la dirección del gesto.

5.4 Obtención de parámetros

Para la creación de cada uno los bloques de movimiento de la etapa de detección de movimiento, se han utilizado 7 puntos como se había mencionado anteriormente, pero para determinar si en un bloque había habido o no movimiento, ha sido necesaria la obtención de un parámetro. La obtención de este parámetro se ha realizado mediante el análisis de las secuencias de puntos correspondientes a los diferentes gestos, para cada mano en particular.
En la siguiente grafica se observa, la trayectoria seguida por una de las manos, durante un movimiento horizontal. En la grafica (5.10) pueden distinguirse claramente las partes correspondientes a un gesto descritas anteriormente, es decir la parte inicial, la correspondiente al movimiento en sí mismo y la parte final.

Se observa:

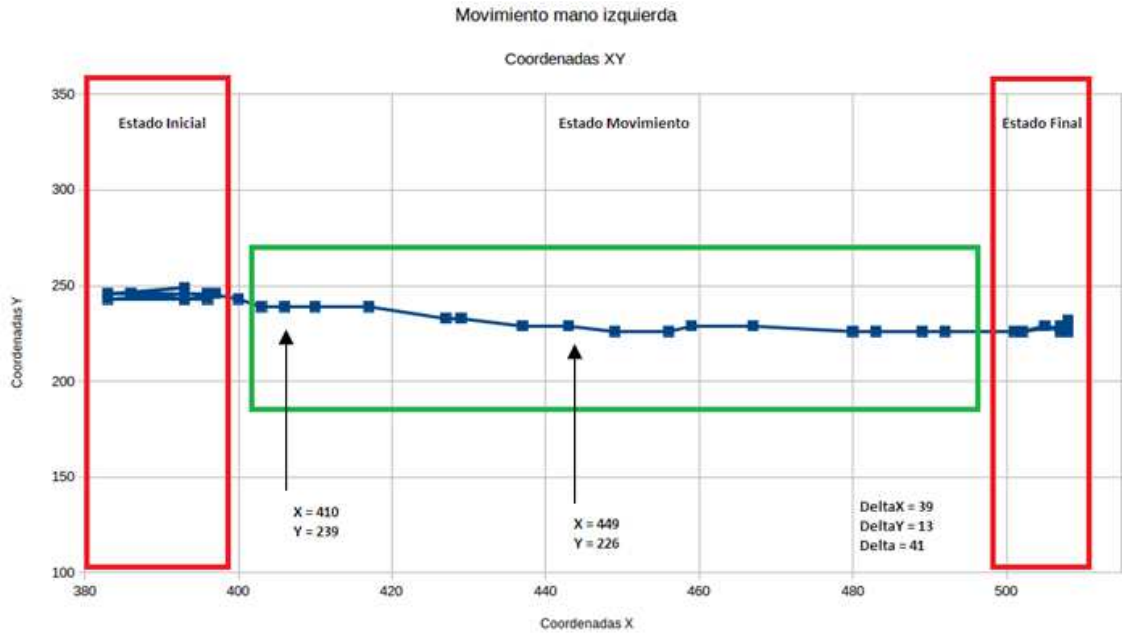

Figura 5.10 Sucesión de puntos generados por el movimiento de la mano izquierda a la izquierda.

Los puntos oscilantes al inicio corresponden al momento en el cual las manos están situadas frente a la cámara, que idealmente deberían permanecer estáticas, pero como el sujeto no puede estar perfectamente estático, los puntos tendrán localizaciones diferentes (hablando en pixeles), eso si dentro de un radio prudencial.

-La parte intermedia corresponde a los instantes concernientes al movimiento, puede verse cómo los puntos tienen una separación mayor entre sí, lo que significa que de un fotograma a otro ha habido una mayor cantidad de movimiento que el que hubo en la parte inicial, además estos se extienden de manera horizontal, como era de esperar ya que el movimiento era de este tipo, también es claro, que esta fase tiene una mayor duración que las demás pues como se ha dicho al principio corresponde al momento efectivo en el que se está realizando el gesto.

-Parte final, es la más pequeña, pero en ella se percibe claramente que el sujeto ha terminado el movimiento, pues los puntos pertenecientes a esta parte de la secuencia, aún a pesar de tener cierta oscilación, ésta es bastante reducida y los puntos se encuentran poco distantes entre sí.

Centrándonos en la parte central de la grafica correspondiente al movimiento, se puede observar

que ya en siete puntos, es decir una secuencia de siete puntos consecutivos, atendiendo a la parte inicial y final de dicha secuencia la distancia Euclidea obtenida con las coordenadas 2D (‘x’ e ‘y’) , ya es de 41, que resulta bastante mayor que si tomáramos la parte inicial y final, de los instantes iníciales o finales de la grafica, pues efectivamente ha habido movimiento, por este motivo se ha tomado este numero de fotogramas para formar los denominados ‘bloques de movimiento’ en secciones anteriores, también como se ha comentado anteriormente debido a la condición acerca de la mínima duración de un gesto, un único bloque de movimiento no es suficiente para considerar que ha habido gesto.

Por otro lado para determinar el umbral a partir del cual se considera que ha habido movimiento, se han tomado diferentes distancias Euclideas de diferentes secuencias de gestos, llegando en un principio a la conclusión, de que en media este valor debía ser igual a 20, sin embargo tras ponerlo a prueba en diferentes secuencias, se ha ido variando para ajustarse mejor a las mismas, de manera que sea válido para la mayoría de ellas, quedando finalmente con un valor de 17.

6- Resultados

Durante la realización del trabajo se ha encontrado inconvenientes como la no detección de las manos, como en el ejemplo de la figura (6.1). Se observa que las manos no se detectan al estar demasiado juntas, pues debido a la resolución de la imagen de distancias, se detecta un único objeto en vez de 2.

Figura 6.1. Fotograma con manos demasiado juntas. Imagen original (arriba izquierda), Imagen distancias (arriba derecha), mínimos de la escena (abajo), se observa cómo las manos aparece como un único objeto y por tanto no son detectables como manos.

Finalmente tras pasar por cada uno de los bloques, se han conseguido los siguientes resultados, hasta finalmente detectar un gesto realizado por el sujeto situado delante de la cámara.

Detección de manos

Figura 6.2. Proceso de detección de manos

A partir de la escena mostrada, se ha generado una imagen de distancias, la cual se utiliza para detectar las manos.

Generación de bloques de movimiento

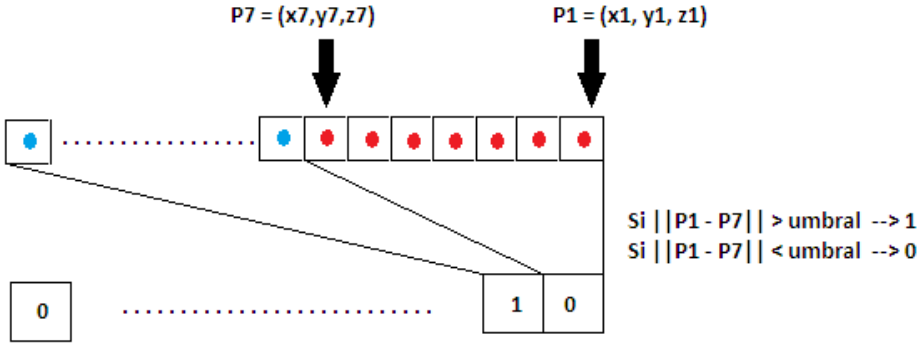

Figura 6.3. Proceso de obtención de bloques de movimiento.

Con las manos detectadas, a partir de sus coordenadas 3D, se generan los llamados bloques de movimiento, que permitirán posteriormente decidir, si ha habido o no gesto.

Detección de gesto mediante Analizador de patrones

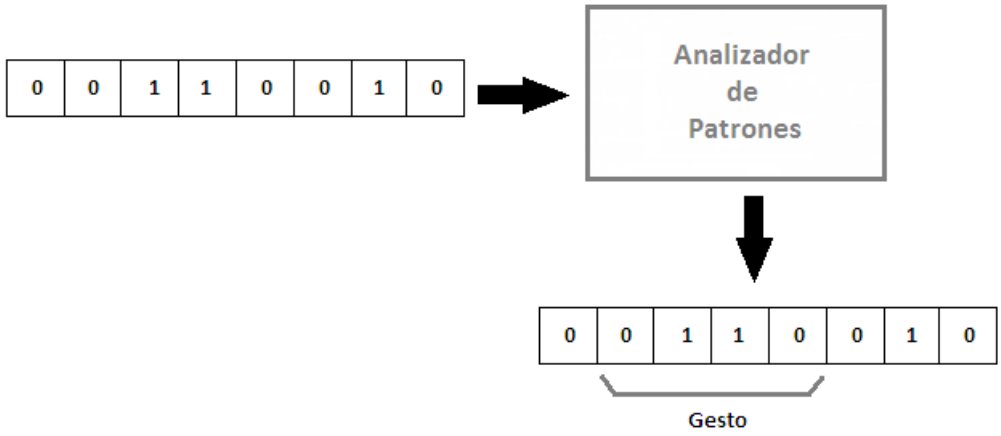

Figura 6.4. Análisis de la secuencia formada por los bloques de movimiento mediante analizador de patrones.

Identificación del gesto

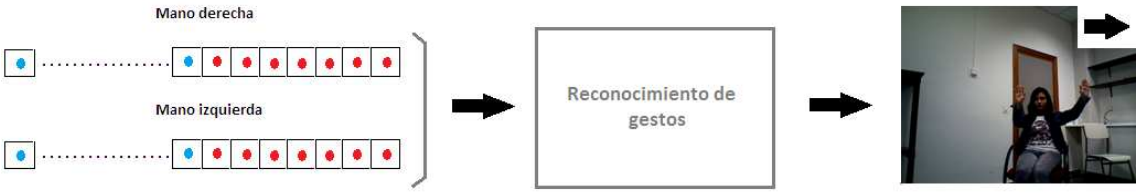

Figura 6.5. Procesamiento de coordenadas de mano izquierda y derecha para identificación de gesto.

Analizando nuevamente las coordenadas de las manos, una vez se ha detectado que ha habido gesto, se debe identificar el tipo gesto realizado, en el caso de la figura (6.5) se trata de un gesto de la mano izquierda horizontal y hacia la izquierda.

Además cabe mencionar las condiciones en las que se produce un buen funcionamiento del programa :

- En primer lugar se ha de detectar que no ha habido movimiento, luego tras una serie de bloques con movimiento se debe volver a detectar otro bloque sin movimiento, momento en el que se detectara el gesto. Aunque este hecho ya fue expuesto en apartados anteriores, el hecho de realizar dos gestos consecutivos sin que haya una mínima pausa entre ambas, provocaría la no detección de dicho gesto.
-Si el movimiento dura más de 2 segundos , no se detectará , pues el analizador no encontrara el patrón, o más bien la secuencia se sobrescribirá con otra más reciente por tanto nunca llegará alcanzarlo.

Finalmente cabe enumerar los resultados obtenidos, es decir los gestos que se ha conseguido detectar, son los mostrados en la siguiente tabla

GESTO		2 MANOS	1 MANO
Arriba		Si	si
Abajo		Si	Si
Agrandar		Si	No
Reducir		Si	No
Izquierda		No	Si
Derecha		No	Si

Tabla 2. Gestos finalmente detectados mediante el programa.

7. Conclusiones

A lo largo de este proyecto se ha pasado por dos etapas, la de detección de mínimos, en la cual se trabaja directamente con la imagen de distancias, y la etapa en la que se trabaja con la información obtenida de la imagen de distancias, correspondiente a los bloques de "Detección de movimiento" y "Reconocimiento de gestos". Debido a que la información procesada es obtenida exclusivamente de la imagen de distancias, se deben tener en cuenta ciertos aspectos al momento tanto de capturarla, como de trabajar con ella:

-El sujeto debe situarse alrededor de 1,5 metros de la cámara.

-Si las manos se encuentran demasiado juntas, aparecerán como un único objeto, por tanto debe haber una separación de mínimamente 10 cm entre los centros de ambas.

-Ha de realizarse un tratamiento previo a la imagen, para eliminar las zonas negras en la imagen. Estas aparecen tanto por las sombras creadas por el propio sujeto, impidiendo que el patrón se proyecte sobre ciertas zonas circundantes, además del tipo de material y color de los objetos en la escena.

Las posibles mejoras que se podrán efectuar al proyecto, son en cuanto a duración y tipo de gesto detectable.
Como se ha mencionado anteriormente la duración de los gestos no debe ser mayor a 2 segundos, pues éste no será detectado, si se deseara detectar gestos con mayor duración, una mejora sencilla sería aumentar este tiempo, modificando el parámetro correspondiente.
En cuanto a los tipos de gestos detectables, se podrá modificar el bloque de "Reconocimiento de gestos", de manera que utilice todos los puntos 3D correspondientes a la mano (o manos) que realiza el gesto para identificar la trayectoria que sigue, además se deberá crear un registro con la trayectoria de los diferentes gestos, para comparar la trayectoria detectada con las del registro.

8.Bibliografía

[1] Dashwood Cinema solutions : A beginners guide to shooting stereoscopic 3D http://www.dashwood3d.com/blog/beginners-guide-to-shooting-stereoscopic-3d/

[2] Problema de correspondencia I:
https://www.cse.unr.edu/~bebis/CS791E/Notes/StereoCorrespondenceProblem.pdf

[3] Problema de correspondencia II:
https://www.google.es/url?sa=t&rct=j&q=&esrc=s&source=web&cd=10&sqi=2&ved=0ahUKEwi4tri5oZrRAhUFNxQKHfX6A5EQFghUMAk&url=http%3A%2F%2Fwww.massey.ac.nz%2F~mjjohnso%2Fnotes%2F59731%2Fpresentations%2FCorrespondence%2520Problem%2520in%2520Stereo%2520Vision.doc&usg=AFQjCNHrm2N4ztTT6SoUd0WDnff9BnM2rg&sig2=Kq4WiRIu5mQs6ojqaSG9pw&cad=rja

[4] Luz estructurada I:
http://web.yonsei.ac.kr/hgjung/Lectures/AUE859/7.%20Structured%20Light%20Projection.pdf

[5] Luz estructurada II:
http://www.sci.utah.edu/~gerig/CS6320-S2013/Materials/CS6320-CV-S2013-StructuredLight-II.pdf

[6] Radu Horaud, Miles Hansard, Georgios Evangelidis, Menier Cl_ement. An Overview of Depth Cameras and Range Scanners Based on Time-of-Flight Technologies. Machine Vision and Applications Journal, 2016. <hal-01325045>

[7] Diapositivas TOF :
http://perception.inrialpes.fr/~Horaud/Courses/pdf/Horaud_3Dcameras_tutorial.pdf

[8] Miles Hansard, Seungkyu Lee, Ouk Choi, Radu Horaud. Time of Flight Cameras: Principles,Methods, and Applications. Springer, pp.95, 2012, SpringerBriefs in Computer Science, ISBN978-1-4471-4658-2. <10.1007/978-1-4471-4658-2>. <hal-00725654>

[9] Sebastian Bauer, Der Technischen Fakultät der Friedrich-Alexander-Universität Erlangen-Nürnberg. *Rigid and Non-Rigid Surface Registration for Range Imaging Applications in Medicine. 2014*
https://opus4.kobv.de/opus4-fau/files/5466/SebastianBauerDissertation.pdf

[10] Imagen de diapositivas asignatura : Tratamiento digital de Imagen .Universidad Politécnica de Valencia. Profesor : Antonio Albiol Colomer.

[11] Richard Hartley and Andrew Zisserman Australian National University, Canberra Australia. *Multiple View Geometry in Computer Vision .ISBN 0521 54051 8 hard back. Second Edition.*

[12] http://um3d.dc.umich.edu/wp-content/uploads/2015/08/Kinect1.jpg

[13] Morfología :
http://iaci.unq.edu.ar/materias/vision/archivos/apuntes/Operaciones%20Morfol%C3%B3gicas%20en%20Im%C3%A1genes%20Binarias%20-%20parte%201.pdf

[14] Segmentación:
http://www.lcc.uma.es/~munozp/documentos/procesamiento_de_imagenes/temas/pi_cap6.pdf

[15] Tyler Bell Beiwen Li Song Zhang. Purdue University, West Lafayette, IN, USA. "*Structured Light Techniques And Applications"*.

[16] Robert Lange. A dissertation submitted to the Department of Electrical Engineering and Computer Science at University of Siegen,*"3D Time of flight distance measurement with custom solid-state image sensors in CMOS/CCD- technology".2000.*
[17] Dong-Ik Ko, Gaurav Agarwal, Texas Instruments. *"Gesture recognition : Enabling natural interactions with electronics"*.

[18] Manuel Armenteros. *"Técnicas audiovisuales : El 3D estereoscópico ha vuelto para quedarse"*.Revista TELOS.

[19]Javier Vargas Balbuena.Universidad Complutense de Madrid. Departamento de óptica. "*Nuevos métodos de medida 3D mediante triangulación activa".2009.*

[20] Milos Davidovic, Michael Hofbauer, Kerstin Schneider-Hornstein, Horst Zimmermann. Vienna University of Technology,*"High Dynamic Range Background Light Suppression for a TOF Distance Measurement Sensor in 180nm CMOS".2011.*

[21] Hamed Sarbolandi, Damien Lefloch, Andreas Kolb. Universität Siegen, *"Kinect Range Sensing: Structured-Light versus Time-of-Flight Kinect".2015.*

[22] S. Burak Gokturk, Hakan Yalcin, Cyrus Bamji, *"A Time-Of-Flight Depth Sensor – System Description, Issues and Solutions"*. Canesta Inc.

[23] David Fofi, Tadeusz Sliwa, Yvon Voisin. Univ. de Bourgogne, *"A comparative survey on invisible structured ligth".2004.*

[24] Leonard McMillan Jr. A dissertation submitted to the faculty of the University of North Carolina, *"An Image-Based Approach to Three-Dimensional Computer Graphics".1997.*

[25]Christoph Mertz, Sanjeev J. Koppal, Solomon Sia, Srinivasa Narasimhan,*"A low-power structured light sensor for outdoor scene reconstruction and dominant material identification"*. Published in Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE
[26] Unai Mujika Torrontegi, Tesis de Máster, Universidad del País Vasco.*"Reconstrucción densa de modelos artificiales utilizando Visión Artificial". 2010.*