UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Métodos objetivos estadísticos de valoración de la magnitud de interés en base a las trayectorias de dirección de visionado

Manuel López Palma

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Departamento de Ingeniería Electrónica

Tesis Doctoral

MÉTODOS OBJETIVOS ESTADÍSTICOS DE VALORACIÓN DE LA MAGNITUD DE INTERÉS EN BASE A LAS TRAYECTORIAS DE DIRECCIÓN DE VISIONADO

Autor: Manuel López Palma

Director: Dr. Javier Gago Barrio
Director: Dr. Montserrat Corbalán Fuertes

Terrassa, septiembre de 2020

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Departamento de Ingeniería Electrónica

Tesis Doctoral

MÉTODOS OBJETIVOS ESTADÍSTICOS DE VALORACIÓN DE LA MAGNITUD DE INTERÉS EN BASE A LAS TRAYECTORIAS DE DIRECCIÓN DE VISIONADO

Autor: Manuel López Palma

Director: Dr. Javier Gago Barrio
Director: Dr. Montserrat Corbalán Fuertes

Terrassa, septiembre de 2020

AGRADECIMIENTOS

Quiero agradecer expresamente la colaboración de los directores Javier Gago y Montse Corbalán sin cuyo apoyo continuo este trabajo no se habría realizado.

Así mismo deseo agradecer el gran soporte dado por Ramon Morros en todo lo relacionado con los algoritmos y librerías de procesado de imagen.

También deseo agradecer todo el cariño y comprensión de mi familia que ha sido muy importante.

ÍNDICE

SIGLAS Y ABREVIATURAS

A

AT, Attention / Engagement time, 13, 27, 28,52

B

Boost Cascade, Impulso en Cascada, 14, 21, 102, 103, 106, 108, 110

C

CDTI, Centro para el Desarrollo Tecnológico Industrial, 13, 26

clicks, 13, 25

D

Dwell time, Tiempo de permanencia, 52, 53

DFOA, Density focus of attention, 7, 8, 14,16, 19, 52, 53, 54, 55, 56, 57, 58, 86, 87, 88, 89, 90, 118, 120, 121, 122, 131, 132

digital signage, señalización digital, 25, 149

E

estimación de la postura de la cabeza, 14

F

FVA, Foco de atención visual, 13

G

GMM, Gaussian Mixture Models, 33

H

HOG, Histogram of Oriented Gradients, 7,14, 92, 102

HMM, Hidden Markov Models, 33

HPE, Head Pose estimation 35

I

IDSP, Intelligent Digital Signage Player, 13, 26

IMU, Inertial Measurement Unit, 13, 16, 17, 19, 29, 42, 82, 85, 86, 118, 124, 125, 127, 133

IVT, In-view time, Tiempo de visualización, 13, 27, 28, 52, 54

L

LBP, Local Binary Pattern, 7, 14,16, 18, 21, 102, 103, 106, 108

LOPD, Ley de protección de datos, 13, 26

M

MFA, Mask Free area, 14, 67

P

PCL, Point Cloud Library, 14, 28, 41, 154

R

RGB, conponentes de color (rojo, verde, azul), 13, 28, 29, 30, 35, 36, 37, 38, 77, 142, 151, 152, 153

RGB-D, componentes de color y depth, 36, 37, 38, 152, 153

S

T

U

V

Y

ÍNDICE DE TABLAS

RESUMEN

En este trabajo se presenta un nuevo método para cuantificar la atención prestada por las personas en los diferentes objetos de su entorno. La nueva métrica modeliza el ojo para determinar la atención prestada y cuantifica dicha atención en todos los puntos de una zona de interés. Se compara esta métrica con la actual basada en tiempo y se justifica que la nuestra se ajusta más a la percepción humana.

Para el cálculo de dicha atención se introduce el concepto de trayectoria orientada como el conjunto de posiciones y de ángulos de orientación de la cabeza, de cada persona de interés y en el tiempo que sea de interés. Justificaremos que solo con estos datos se puede determinar dicha atención.

En el método presentado se utilizan cámaras cenitales como sistema de tener las mayores prestaciones con el mínimo número de cámaras. Así mismo se analizan dos métodos: el método 3D que utiliza la información de profundidad, y un método menos preciso, el 2D, que solo utiliza cámaras de imagen.

Esta tesis también presenta una forma de calcular la métrica de tiempos, método que se utiliza ampliamente en la actualidad para verificar cuantas personas y en cuánto tiempo se ha visto un anuncio La forma presentada por nuestro método permite reducir el número de cámaras necesarias, y por tanto es ventajosa en cuanto a los recursos que requiere para su implementación.

Finalmente se verifican los resultados utilizando una cámara en la parte frontal de la cabeza simulando al ojo y un sensor IMU que mide los ángulos de la cabeza. De esta manera se determina la relación de atención de los objetos detectados por la cámara, y la misma relación de atención de los objetos obtenidos por el método propuesto.

SUMMARY

In this work, a new method is presented to quantify the attention paid by people to the different objects in their environment. The new metric models the eye to determine the attention given and quantifies that attention at all points in an area of interest. This metric is compared with the current one based on time and it is justified that ours is more in line with human perception.

For the calculation of said attention, the concept of oriented trajectory is introduced as the set of positions and orientation angles of the head, of each person of interest and in the time that is of interest. We will justify that only with this data can such care be determined.

In the presented method, top view cameras are used as a system to have the highest performance with the minimum number of cameras. Likewise, two methods are analyzed: the 3D method that uses depth information, and a less precise method, 2D, that only uses imaging cameras.

This thesis also presents a way of calculating the time metric, a method that is widely used today to verify how many people and in how long an ad has been seen. The form presented by our method allows reducing the number of cameras required, and therefore it is advantageous in terms of the resources required for its implementation.

Finally, the results are verified using a camera in the front part of the head simulating the eye and an IMU sensor that measures the angles of the head. In this way, the attention relationship of the objects detected by the camera is determined, and the same attention relationship of the objects obtained by the proposed method.

1 INTRODUCCIÓN

Disponer de un método para determinar a que prestan atención las personas es algo muy buscado desde hace algún tiempo. Sectores como el publicitario, el márquetin requieren de sistemas que puedan determinar que productos o que anuncios son los que crean una mayor atracción a las personas. Esta necesidad se agudiza con la llegada de internet donde se traza el comportamiento humano a partir de las búsquedas y acciones que se realizan en la web. Sin embargo, cuando las personas se mueven por recintos comerciales; no existe en la actualidad una forma bien establecida para poder asociar las preferencias; o mejor poder comparar que productos o señales llaman la atención de las personas.

En el presente trabajo propondremos una métrica que tendrá como objetivo determinar tanto que objetos tienen mayor interés; en base a la atención prestada por el individuo.

Determinaremos el enfoque de atención (FoA) de las personas en un espacio cerrado (por ejemplo, un área comercial), y expondremos como esta métrica puede tener gran interés para múltiples aplicaciones como la tarificación publicitaria, seguridad, ventas al por menor, en la gestión de marketing de locales comerciales y como instrumento de valoración de flujo humano en locales y lugares públicos.

Además, también relacionaremos el Foco de atención, con una generalización del concepto de Tiempo de atención [1]. Esta medida cuantifica la cantidad de atención que recibe una región durante un período de tiempo, pero la nueva métrica presentada tiene en cuenta más factores como la distancia al objetivo, la velocidad de movimiento del espectador y el ángulo de visión con respecto a la trayectoria. De esta manera, se puede proporcionar una determinación más realista de la atención cuando se mueve en un área.

1.1 MOTIVACIÓN

La presente tesis surge como evolución de los trabajos realizados en la empresa Venco para el contaje de las personas que están viendo un anuncio (los impactos que ha tenido dicho anuncio) y por tanto para el cambio de tarificación de la publicidad, que pasa de

tarificar basándose en el número de reproducciones a basarse en cuanta gente ve el anuncio (impactos).

En el mercado del *digital signage* estos tipos de medidas tuvieron y tienen un gran auge porque permiten comparar el resultado de la publicidad de una forma similar a los *clicks* que se realizan en los anuncios en internet.

En Venco la empresa en que el doctorando trabajaba como director de I+D se habían realizados diseños y comercializados productos que cuentan las personas; caracterizándolas con género y edad, que ven un determinado anuncio. El anuncio se realizaba sobre pantallas publicitarias y utilizando cámaras de visión frontal. El doctorando ha dirigido los proyectos relativos al tema anterior que se mencionan seguidamente:

- Dirección CDTI en la empresa Venco "Sistema de Evaluación de la audiencia en las pantallas publicitarias de interior; presentado y aprobado por CDTI con número de expediente IDI-20101193".

- Codirección proyecto Eurostar formado por dos empresas Trumedia (Israel) y Venco (España) "Intelligent Digital Signage Player IDSP" número de expediente E! 5644 –IDSP.

La técnica que utilizábamos era la de detectar las personas a través de la cara, mediante una cámara en el punto que se desea computar dicha atención, normalmente es la pantalla donde se emite el anuncio. El problema que tiene este método es que para cada punto de análisis requiere una cámara y, por tanto, como evolución natural se planteó, se pidió y se concedió un tercer CDTI, donde el objetivo era la detección de los impactos de toda una zona; pero con un número reducido de cámaras.

Sin embargo, cuando empezamos a trabajar con este tema nos dimos cuenta de que dicho problema no había estado estudiado y fue imposible encontrar una solución. Entonces la empresa Venco tuvo reuniones con los que acabaron siendo los directores de la Tesis con el objeto de establecer una colaboración para encontrar un método implementable en un producto que pudiese superar esta barrera. De esta forma se obtuvo financiación para realizar el Doctorado Industrial mediante un proyecto de la AGAUR (Generalitat de Catalunya) con referencia 2014DI067, y se firmó un convenio de colaboración con la UPC.

Nuestro objetivo se basa en capturar la escena con una cámara cenital y determinar el enfoque de atención de una persona utilizando su trayectoria y la dirección de la cabeza. Si bien las soluciones basadas en el seguimiento ocular podrían determinar más exactamente la dirección de la mirada, son complicadas en escenarios grandes o complejos, y requieren una cámara de alta resolución en cada ubicación en la que se debe medir la atención. Nosotros pretendemos determinar la atención en cualquier punto con una sola (o unas pocas) cámaras.

La configuración de vista superior presenta otras ventajas, como ser casi inmune a las oclusiones, evitando la captura de la cara, con lo que eliminaremos dificultades de grabación de imágenes asociadas a la Ley de protección de datos LOPD [2]. Además, se puede obtener información adicional, como la distancia de visualización y el ángulo relativo de visualización, lo que permite un análisis más completo de la escena.

La presente tesis industrial pretende realizar una implementación que demuestre a la empresa que el método propuesto es realizable.

1.2 OBJETIVOS

El objetivo de la tesis es encontrar un método que permita evaluar la atención prestada, por todas las personas que pasen por una zona (zona de análisis). El método debe utilizar un número reducido de cámaras en comparación con los métodos utilizados anteriormente por Venco y debe de poder extraer resultados equivalentes a los métodos de cámaras de visión frontal con las que se calculan las métricas de medición de tiempo *In-view time* (IVT) y *Attention / Engagement time* (AT) [2]. La IVT mide la cantidad total de tiempo que el observador se encuentra, por ejemplo, en la zona de visibilidad del anuncio (no necesariamente prestándole atención), mientras que AT es la cantidad total de tiempo que el observador está observando activamente el anuncio. AT permite cuantificar el grado de atención que ha recibido un anuncio dado.

Sin embargo, nuestro punto principal es buscar otra métrica más justa que la actual. De hecho, la métrica actual que utiliza la acumulación de tiempo para indicar cual ha sido la atención en una zona adolece de un punto principal. El tiempo tiene el mismo valor si la persona está cerca o lejos del anuncio; pero el impacto que produce sobre la persona no es

el mismo, dado que será mayor cuanto más cerca esté del anuncio. Por tanto, nuestro objetivo será establecer una métrica que tenga en cuenta este hecho y permita distinguir la atención recibida teniendo en cuenta el efecto que produce sobre el cerebro humano.

Por otra parte, deberíamos de poder valorar de forma tanto subjetiva como objetiva el resultado del método.

En base a ese objetivo principal, se proponen los siguientes objetivos parciales:

- creación y justificación de una métrica o método de determinación de la atención visual
- Seleccionar e instalar el *setup* adecuado para la captación de los vídeos
- Generación de los vídeos para la validación del método propuesto
- Selección del algoritmo para la implementación del método
- verificación del método
- Confirmación que el método empleado es implementable en un producto
- Utilización de software y algoritmos libres OpenCV, Dlib, PCL,.. Python, C++…

1.3 PLANTEAMIENTO DEL PROBLEMA

Como hemos indicado previamente los métodos del contaje de personas de Venco utilizaban exclusivamente cámaras para la detección de las caras. El nuevo método no podía utilizar los mismo porque tendría la misma problemática, un numero de cámaras muy elevado. Planteamos una aproximación distinta; en lugar de analizar donde está la cara intentaremos predecir hacia donde mira la persona, pero utilizando métodos indirectos. Para ello utilizaremos la detección del cuerpo que nos permita identificar hacia donde presta atención. En el inicio hicimos la hipótesis de que dado que los ojos están en la cabeza solo es necesario caracterizar cual es la posición y orientación de esta. A partir de aquí se planteó como cubrir el espacio más amplio posible con el menor número de cámaras permitiendo indicar con detalle las cabezas existentes y sus desplazamientos. Inicialmente utilizamos

cámaras de imagen RGB; pero con la evolución del proyecto nos encontramos que solo con cámaras de imagen el problema no tiene solución precisa, y evolucionamos a cámaras RGB+D (3D).

El problema por tanto consiste:

- Obtención de un modelo que permita determinar qué atención presta una persona cuando su cabeza está posicionada en un punto del espacio y tiene una determinada orientación.
- Obtención de la formulación que calcule tanto los tiempos (IVT y AT) como el modelo de atención descrito en el punto anterior, a partir de las posiciones de la cabeza de un conjunto de individuos en una zona de análisis y durante un periodo de tiempo.
- Implementación de una solución que sea operativa para convertir en un producto. Lo que hagamos debe de implementarse en soluciones de un coste razonable. Adicionalmente debe de respetar las normas de confidencialidad existentes.
- Verificación del método. Hemos de concebir un sistema que sin ninguna duda convenza a cualquier persona que el método planteado coincide con lo que la persona ha visto durante su paso.

Para ser útil, la aplicación debe ser robusta, no invasiva y adaptable a diferentes entornos; a menudo llena de muchos objetos. Otro requisito importante es que la configuración completa debe ser lo más barata posible para que sea competitiva en una variedad de situaciones. La razón es permitir que el sistema se ejecutara en hardware simple, sin la necesidad de una GPU que aumentaría el costo. Por esta razón se descarta el uso de algoritmos de aprendizaje profundo para la detección de la cabeza, la estimación de ángulo y del seguimiento, en esta primera fase.

1.4 MEDIOS EMPLEADOS

Para la implementación se han usado medios materiales y medios humanos. Con respecto a los medios humanos se ha contado con el doctorando con el soporte de los directores de la tesis y con el soporte de Ramón Morros del Departamento de Teoría de la Señal y

Comunicaciones. Adicionalmente se han realizado siete Trabajos final de Grado asociados a esta Tesis.

Con respecto a los medios materiales; se han realizado pruebas en dos instalaciones en la empresa Venco y en la UPC de Terrassa, inicialmente se partió de una red de 5 cámaras RGB cenitales con una estructura cuadrada con una cámara en el centro y posteriormente se pasó a una cámara RGB+ Depth (3D) también cenital. Además, se ha contado con una cámara RGB y con una unidad IMU solidarias a la cara en los ensayos de validación de pruebas 3D. Para la validad se ha usado también una cámara RGB en la pared para simular los métodos actuales.

Para el procesado y almacenamiento se han utilizado ordenadores personales core i7 32GB dram y con discos SSD de 2TB; si bien se han utilizado otros centros de cálculo de la UPC para el trabajo de verificación de redes neuronales convoluciones. Con respecto a la sincronización de capturas puesto que existían diferentes procesadores leyendo cada cámara se utilizó el sistema ROS que permite conjuntar los datos de forma sincronizada.

Los ensayos se realizaron con personal de la UPC (profesores, PAS y estudiantes) y también con personal de la empresa Venco.

1.5 ESTRUCTURA DE LA MEMORIA

Para facilitar la lectura de la memoria, se incluye a continuación los capítulos en los que hemos divido la Tesis. En general la memoria está separada en métodos 3D y 2D dependiendo de si la cámara utilizada es la RGB+D o solo la cámara RGB. En el primer caso dispondríamos de la información de la profundidad para determinar el algoritmo.

El primer capítulo realizamos la introducción de la tesis y exponemos los objetivos.

El segundo capítulo es una revisión del estado del arte sobre los métodos utilizados hasta la fecha y una bibliografía sobre toda la temática trabajada en esta tesis.

En el tercer capítulo se describe y justifica el método. Este es el capítulo fundamental puesto que es realmente lo que aportamos como novedad, los demás capítulos solo presentan

métodos de realizar el cálculo, pero seguro cambiaran con el paso del tiempo, pero el método encontrado esperamos que permanezca en el futuro. Este método se diferencia según si partimos de la información 3D o 2D; el método 3D es el que consideramos válido; mientras que el método 2D es una aproximación que solo puede utilizarse para sistemas de bajo coste donde se acepte una aproximación al cálculo de atención.

En el cuarto capítulo se describe la implementación, donde se presentan los diferentes métodos que hemos utilizado para determinar las posiciones de la cabeza y ángulos. Al igual que en el caso anterior separamos las dos implementaciones 3D y 2D.

La verificación de los resultados, también separando el caso 3D y 2D, se exponen en el capítulo cinco.

Y, por último, mostramos el capítulo de conclusiones y líneas de trabajo futuras.

2 ESTADO DEL ARTE

En el estado actual existen diferentes algoritmos para la identificación de personas u objetos. La mayoría de las referencias detectan objetos estáticos. En [3] reconocen caras, personas y coches estáticos. Presentan un sistema de clasificación y reconocimiento de patrones con porcentajes de acierto altos. Algunas técnicas están orientadas a extraer la silueta perfecta lo cual puede ser complicado en función de la escena, debido a que presente variaciones de iluminación, un suelo que dificulte el reconocimiento de las personas por problema de contraste o por ejemplo confusión con sombras [4]. Otras técnicas se basan en la detección de regiones de la persona o características y en aplicar modelos de clasificación sobre ellas o aprendizaje supervisado [5]. Estos sistemas no necesitan la obtención de una forma muy definida, sino que con una buena aproximación es suficiente. En [6] utilizan para la detección de las personas cuatro detectores, uno para: la cabeza, las piernas, el brazo izquierdo y el brazo derecho. Manejan una arquitectura de clasificación jerárquica, en la que el aprendizaje ocurre en más de dos niveles. Presentan resultados que muestran que su sistema tiene un rendimiento significativamente mejor que un detector de personas de cuerpo completo. Su sistema es más robusto para localizar vistas parciales de personas y personas cuyas partes del cuerpo tienen poco contraste con el fondo.

En general tiene mayor interés la detección de objetos o personas en movimiento utilizando secuencias de video [7][8][9][10][11]. Algunos métodos se basan en la segmentación del frente de la imagen del fondo. Al extraer el fondo (background) se obtiene una imagen binaria con los píxeles del frente de la imagen original (foreground) blancos y los píxeles del fondo de la imagen original (background) negros. Esta imagen binaria se divide en segmentos (blobs) que están constituidos por un conjunto de píxeles conectados y que representan a los individuos o grupos de personas. Para cada blob se generan características diferentes, por ejemplo, en [12] utilizan la generación del histograma de color, se calcula la orientación y la longitud del borde.

Según la aplicación lo que interesa es detectar a la persona y saber trazar su trayectoria [11][13][10][14][15][16][17]. Los autores P. Dollar et al. [16] realizan una extensa evaluación del estado del arte en un marco unificado sobre el número de enfoques que hay para detectar peatones en imágenes monoculares. En [18] pretenden detectar a los peatones

y modelizar su comportamiento para predecir posibles colisiones entre peatones y vehículos. En otros trabajos se quiere contabilizar el número de personas que pasan a través de una posición concreta [19][20]. Detectar y rastrear personas es un requisito clave en el desar*roll*o de tecnologías robóticas destinadas a operar en entornos humanos. En entornos atestados, como las estaciones de tren, esta tarea es particularmente desafiante debido a la gran cantidad de objetivos y las frecuentes oclusiones. En [21] se muestra un marco para detectar y rastrear a las personas en entornos atestados. Las principales contribuciones son un método para extraer la postura de las partes más visibles del cuerpo en una multitud, la cabeza y los hombros, y un rastreador que aprovecha las limitaciones sociales relacionadas con la orientación, el movimiento y la proximidad de las personas, para mejorar la robustez en este ambiente desafiante. Consiguen logros, aunque también les aparecen restricciones que deben replantearse en el futuro.

Para detectar a personas otros investigadores han optado por detectar la cara de las personas en ocasiones para lograr los mismos objetivos, Chen et al. [20] cuentan la gente que pasa a través de una puerta detectando las caras. La tecnología de seguimiento ocular se ha utilizado para analizar la dirección de la mirada del cliente y determinar si él o ella está mirando activamente a un determinado producto o señal. Por ejemplo, [22] investigan la prominencia visual en una tienda de sus productos y cómo ésta prominencia afecta a las decisiones del cliente. El análisis se realiza mediante el uso de datos de seguimiento visual y datos de ventas en la tienda de comestibles. También se utiliza el seguimiento ocular. Las cámaras frontales pueden ubicarse en o cerca del producto o señal para tener una vista frontal de la persona. En esta posición frontal, la cara y los ojos de las personas pueden ser detectados, haciendo posible un análisis fino de la dirección de la mirada. Las vallas publicitarias y los letreros se han aplicado de manera amplia y exitosa para enviar información y publicidad [23]. Sin embargo, cómo evaluar los beneficios enviados es un tema extremadamente importante para los dueños de negocios. Para abordar este problema, este documento propone un sistema de evaluación eficaz para las audiencias de cartelera digital. A través de la cámara frontal incorporada, se pueden registrar con precisión las fechas observadas, los periodos de tiempo, las duraciones interrumpidas y las ubicaciones de las audiencias individuales, y a través de estos datos recopilados, el sistema de evaluación propuesto puede proporcionar información estadística. En [24] se estudia la atención que los transeúntes prestan a un anuncio al aire libre usando una sola cámara de video. Se analiza el

foco de atención visual (VFOA) para un número variable de personas (logran hasta tres) sin movimiento restringido. Su método consta de dos componentes: una red bayesiana dinámica, que rastrea simultáneamente personas en la escena y estima su pose de cabeza, y dos modelos VFOA para múltiples personas basados en modelos de mezcla gaussiana (GMM) y modelos ocultos de Markov (HMM), que infieren VFOA de un sujeto desde su ubicación y pose de cabeza. En [25] se investiga el papel y las limitaciones de la visión periférica para las tareas de elección basadas en preferencias en un entorno real como es un supermercado. Drouard et al. [26] and Kuhnke et al. [27] estudian la estimación de la posición de la cabeza. En [26] el método de mapeo combina múltiples técnicas de aprendizaje no supervisado y de mezclas de regresiones que aprende a mapear vectores de características de alta dimensión (extraídos de los cuadros delimitadores de caras) en el espacio conjunto de los ángulos de la postura de la cabeza. En [27] proponen un método para que la información de la postura de la cabeza se utilice para producir mejores alineaciones faciales para el reconocimiento de expresiones o rostros invariantes en las posturas. Se trata de un enfoque actual basado en el aprendizaje profundo supervisado.

La señalización digital adaptativa de la audiencia es una nueva tecnología emergente, en la que las pantallas de difusión pública adaptan su contenido a las características demográficas y temporales de la audiencia. En [28], se presenta un estudio de la medición de la audiencia de señalización digital utilizando cámaras ubicadas en las pantallas de señalización y frente a los clientes. Las métricas temporales del tiempo de permanencia de una persona, el tiempo de visualización en pantalla y el tiempo de atención se extraen por la detección del cuerpo y del rostro y por la estimación de posturas. También se obtiene una estimación del género y la edad de los clientes. En [29] presenta un nuevo enfoque para el modelado automático del comportamiento del consumidor en una tienda minorista de ropa basado en datos de medición de la audiencia. Entre otros parámetros, se han utilizado los tiempos de *In-view* y *Attention*. Muestran que, en un entorno no controlado, los datos de la audiencia se pueden usar para predecir las decisiones de compra. En el trabajo de R. Ravnik y F. Solina [30] usando una cámara monocular instalada dentro del marco de una pantalla de señalización digital, extraen las características temporales, espaciales y demográficas de los observadores. También extraen el número de observadores, su distancia a la pantalla midiendo la distancia interpupilar de las caras registradas, su género y su edad. Lo prueban a nivel de laboratorio y en un entorno real. El objetivo de este número especial [31] es ofrecer

una descripción general del estado de la técnica y los desarrollos recientes en métodos y sistemas para mediciones de audiencia en minoristas y señalización digital. En [32] se presenta la utilización de un sistema de cámara frontal para rastrear las manos de un usuario para fines de reconocimiento de gestos hechos con las manos. La distracción y la falta de atención del conductor son causas importantes de colisiones por eso en [33] estudia el foco de atención de un conductor. Para permitir que los sistemas de asistencia al conductor aborden estos problemas se presenta un nuevo sistema que consta de tres módulos interconectados que detectan la cabeza del conductor, proporcionan estimaciones iniciales de la postura de la cabeza y siguen continuamente su posición y orientación en seis grados de libertad. Por ejemplo, [34] investiga la relevancia visual de la señalización y los productos en la tienda y cómo esta relevancia afecta a las decisiones del cliente. El análisis se realiza utilizando datos de seguimiento visual y datos de ventas de las tiendas de comestibles. Borji y Itti [35] hacen una revisión sobre los avances recientes en modelado de atención visual con énfasis en la prominencia de abajo hacia arriba. Se revisan muchos trabajos utilizando la comparación cualitativa de más de 13 criterios experimentales. Llegan a la conclusión que se está resolviendo problemas de visión desafiantes, como la interpretación de escenas y el reconocimiento de objetos. Sin embargo, aún continúan habiendo muchos problemas sin resolver.

En otras ocasiones es suficiente con el control de una parte del cuerpo humano para poder lograr los objetivos. En [36] se quiere realizar el seguimiento de la atención en reuniones. Proponen un sistema capaz de estimar el enfoque de atención de los participantes a partir de múltiples señales. En el sistema, emplea una cámara omnidireccional para rastrear simultáneamente los rostros de los participantes sentados alrededor de una mesa de reuniones y utiliza redes neuronales para estimar sus posturas. Además, usan micrófonos para detectar quién está hablando. El sistema predice el enfoque de la atención de los participantes a partir de la información acústica y visual por separado, y luego combina la salida de los predictores de atención basados en audio y video. Además, analiza cómo de bien podemos predecir el foco de atención de un sujeto únicamente en función de la orientación de su cabeza, obteniendo en un 89% del tiempo una correspondencia. En este campo hay varios trabajos en los que cambia los elementos estudiados en el enfoque de atención visual, si se analizan a las personas individualmente o en conjunto y las partes del cuerpo que se utilizan [37] [38] [39][40]. En la literatura hay una buena cantidad de trabajos

que utilizan visión por computadora y otras tecnologías de detección para analizar la atención que la gente presta a los carteles públicos. Una buena revisión se puede encontrar en [41] en donde en general se requieren una cámara en cada punto analizado. Una cámara colocada sobre el anuncio de cara al cliente puede medir si un cliente está activamente mirando hacia el anuncio o no. El inconveniente de estos métodos es que requieren una cámara para cada punto de medición, que puede resultar engorroso y afectado por oclusiones. Recientemente se ha publicado el trabajo [42] en donde la estimación de la mirada requiere realizar la estimación de la postura de la cabeza (HPE) basado en el seguimiento de puntos faciales 2D, estimando la postura incremental de la cabeza que da cuenta de la variación en la observación de la imagen en su trayectoria. Además, utiliza una cámara 3D para poder evaluar los ángulos pitch (ángulo de inclinación), yaw (ángulo de giro horizontal) y roll (ángulo de giro vertical). Este trabajo se asemeja a lo propuesto en este trabajo a excepción que nosotros no utilizamos en ningún momento el rostro de las personas.

Los dispositivos móviles son equipos esenciales hoy en día. Euclides et al. [43] proponen un método de estimación de pose de la cabeza en tiempo real para construir una aplicación que permita nuevas formas de interacción humano-móvil. Para captar las imágenes utilizan la cámara frontal del móvil, por tanto, se capta la cara. Referente al movimiento de la cabeza tienen en cuenta tres grados de libertad porque consideran el movimiento *pitch*, el *yaw* y el *roll*.

El inconveniente que tiene la configuración de la cámara frontal es que normalmente requiere de una cámara en cada posición de análisis y está afectada por oclusiones. Por este motivo otras de las configuraciones utilizadas es colocar la cámara en el techo. La vista superior es un método no invasivo y puede evitar los problemas de oclusión de las cámaras frontales. Además, es una solución más rentable porque con una sola cámara se puede analizar varios puntos de interés. El inconveniente de no capturar la cara y los ojos de las personas es que se tiene poco menos de precisión cuando se evalúa la dirección de la mirada. La disponibilidad de la información de profundidad ha sido una forma de mejorar los resultados de detección de personas. La adopción de una configuración de vista cenital reduce la complejidad del problema, al mismo tiempo que permite preservar la privacidad de las personas analizadas. Se ha usado las cámaras en posición cenital para contar personas en [44] y [45]. Concretamente en [45] se combina la utilización de una cámara RGB con una de profundidad y demuestran que el rendimiento se mejora con la adopción combinada de

ambos sensores, lo que permite que el sistema logre una compensación más equilibrada entre falsos positivos y falsos negativos. En [46] se realizó la detección, seguimiento y análisis de patrones de comportamiento de las personas que cruzaban el campo de visión de la cámara RGBD para contarlas. Para ello utilizaron dos técnicas propias del campo de análisis de imágenes 2D trasladadas al contexto de imágenes de profundidad. En [47] se propone un método para la detección humana, seguimiento y contaje consistente en varios pasos algorítmicos. En el primero buscan en el mapa de profundidad los máximos locales. Estos se utilizan como puntos de partida para un algoritmo de localización de cabezas en el siguiente paso. Los centros de las cabezas candidatas se encuentran en este paso y se entregan al detector principal basado en clasificación. El detector de cabeza predice si el candidato principal es una cabeza válida. En un paso final, un algoritmo de seguimiento agrega las detecciones a lo largo del tiempo y calcula las trayectorias para cada cabeza.

Las cámaras RGB-D son una opción en muchos trabajos [48], [49], [50], [51], [52] debido a la capacidad de capturar profundidad. La información adicional a RGB simplifica significativamente la segmentación de los cuerpos y extremidades de las personas, lo que permite un análisis más preciso y potente. En [48], además de una amplia variedad de cámaras RGB-D, se utilizan otras modalidades de sensores (balizas de radio activas emitidas por los dispositivos móviles de los clientes) para determinar la localización del cliente en la tienda. Esto permite capturar información de las actividades del consumidor en la tienda. Un uso popular de los sensores RGB-D es colocarlos en el techo. Por ejemplo, en [49] una cámara RGB-D está situada encima de un estante. Se analizan varios comportamientos de los clientes, incluido el gesto de alcanzar objetos en la estantería, la navegación y el pesaje de productos, etc. Su método determina qué estante alcanza el cliente. Un enfoque relacionado con éste artículo es el [50] donde se propone un sistema de reconocimiento de la postura y la actividad humana en tiempo real, con una sola cámara de detección de profundidad de vista superior. Este método es capaz de rastrear las posiciones y orientaciones de los usuarios, así como reconocer posturas y actividades (de pie, sentado, señalando, etc.). En [52] un conjunto de algoritmos de visión computarizada, incrustados en las cámaras RGB-D distribuidas, proporciona información sobre el comportamiento del cliente, en particular, las interacciones entre usuarios y estantes descritas con datos temporales y características espaciales. La viabilidad y la eficacia de la arquitectura y el enfoque propuestos se han probado en entornos minoristas reales. Los autores indican que

la información recopilada se puede utilizar para varios análisis estadísticos útiles, ya que mejoran, por ejemplo, el conocimiento de las interacciones entre los compradores y los estantes y el objeto deseado.

En [53] también proponen un método para detectar personas usando imágenes con profundidad de una cámara colgada en el techo. En este artículo mejoran un método propuesto en 2013 de forma que ahora son capaces de detectar cabezas que no están enteras. Zhang et al. [54] proponen usar el llamado método "water filling" para detectar y contar gente usando una Kinect con visión vertical. Dicho algoritmo simula el proceso de las gotas de lluvia que caen del cielo a la tierra y se mueven a la región hueca más cercana. Toman el mapa de profundidad como una tierra y arrojan la lluvia sobre él, la región reunida con gotas se clasificará como hueca, lo que significa la cabeza de las personas porque es además la más cercana a la Kinect. Muhammad et al. [55] son capaces de detectar personas medio ocultas y varias personas en una escena a partir de imágenes de profundidad captadas lateralmente de forma que se vean las personas de los pies a la cabeza.

Otros autores también han utilizado la Kinect. En [51][56] se presenta un sistema para la detección de la cabeza en tiempo real y la estimación de la postura de la cabeza a partir de datos de profundidad de baja calidad capturados con un sensor Kinect. Utilizan un bosque de regresión aleatoria adoptando generalmente la curvatura media y gaussiana como funciones de imagen complementaria para el entrenamiento del bosque aleatorio, que clasifica los parches de imagen de profundidad entre la cabeza y el resto del cuerpo y que realiza una regresión en los espacios continuos de las posiciones y orientaciones de la cabeza. Los árboles que conforman el bosque están capacitados para optimizar conjuntamente su poder de clasificación y regresión maximizando dos medidas separadas. En [57] hacen la estimación de la postura de la cabeza a partir de los datos RGB-D por medio de un enfoque de seguimiento temporal de cuadro a cuadro, que incorpora la información temporal para estimar la postura de la cabeza a lo largo de la secuencia de video. Está inspirado en un objeto de seguimiento temporal recientemente propuesto [58],[59] que utiliza imágenes de profundidad y bosques de regresión para estimar la pose de un objeto de un cuadro a otro. La diferencia es que en ese trabajo abordan el problema de adaptar y generalizar el rastreador a las variaciones específicas de cada sujeto, incluso levemente no rígidas, de las estructuras de la cabeza con respecto a un modelo 3D predefinido de la cabeza. Para una aplicación real, hay varios algoritmos de seguimiento para cámaras cenitales (ver, por ejemplo, [60] para

una revisión de los métodos) con un rendimiento excelente que se puede usar para este propósito. En particular, utilizando un método de filtro de partículas es posible rastrear simultáneamente la posición y el ángulo de la cabeza. Aunque hemos optado por otra solución diferente para lograr los mismos objetivos.

El análisis del comportamiento humano a través de la información visual ha sido un tema de investigación muy activo en la comunidad de la visión artificial. En este trabajo [61] se utilizan varias cámaras de profundidad para mejorar significativamente la calidad de seguimiento y reducir las ambigüedades, como, por ejemplo, las oclusiones. Al fusionar las imágenes de profundidad de todas las cámaras pueden estimar mejor la postura. Logran rastrear con precisión el movimiento humano en tiempo real (15Hz) en una GPU. En particular, en [62] se revisan las principales investigaciones publicadas sobre el uso de imágenes de profundidad para analizar la actividad humana. La revisión exhaustiva aborda el modelado corporal 3D articulado para la estimación de posturas humanas y el reconocimiento de acciones humanas. Por ejemplo, en el análisis de eventos sociales, la información de la postura de la cabeza en 3D ayuda drásticamente a determinar la interacción entre las personas y extraer el foco visual de atención [63]. El reconocimiento de actividad personal puede realizarse estimando el foco visual de atención de una persona. Teniendo en cuenta el conjunto de escenarios de sentado en una silla fija donde solo se ve la parte superior del cuerpo, en [64] se enfocan en la cabeza de la persona como una pista importante para el enfoque visual de la estimación de la atención usando una sola cámara RGB-D frente a la persona y proponen una extensión del método de estimación de la postura de la cabeza de [65].

Este articulo [66] presenta un sistema para la detección de personas, la segmentación y el seguimiento de esqueletos para los datos RGB-D que es oportuno para los robots móviles de interiores y muchas otras aplicaciones. Brutti y Lanz [67] tienen en cuenta la información audio visual para determinar la orientación de la cabeza, el desplazamiento de múltiples personas y la actividad de hablar entre ellos. Utilizan un filtro de partículas multimodal. Detecta automáticamente a las personas a medida que ingresan en la sala monitoreada, adquiere su firma visual y rastrea su posición, orientación de la cabeza y actividad del habla utilizando las probabilidades audiovisuales y la técnica de fusión que proponen. Los autores de [68] presentan un enfoque para el seguimiento robusto de rostros en tiempo real utilizando la estimación de la postura de la cabeza para un sistema de realidad aumentada sin

marcadores. Cuando falla el algoritmo de seguimiento, utiliza una estimación de la posición de la cabeza para dar una estimación inicial del algoritmo del punto más cercano iterativo que usan. Esta es la novedad que presentan en el algoritmo de seguimiento de la Kinect Fusion respecto a trabajos anteriores. Además, la estimación de la pose de la cabeza es adecuada para aplicaciones de realidad aumentada, ya que se ejecuta en tiempo real. Estos tres trabajos se alejan de las propuestas de la tesis, pero los hemos querido mencionar como ejemplo de la amplitud de aplicaciones que tiene saber medir la postura de la cabeza.

3 MÉTODO

3.1 INTRODUCCIÓN

Como ya se ha mencionado el objetivo es poder medir cuantitativamente en qué se fija la gente cuando transita por un espacio. Esto es importante en diferentes campos, por ejemplo, en publicidad para saber qué anuncios despiertan más interés. Poderlo determinar mediante un sistema competitivo para la industria es nuestro segundo objetivo. Por ello apostamos por la utilización de las cámaras en la posición cenital en contra de usar una cámara para cada elemento susceptible de llamar la atención y que se está analizando.

Concretamente se utiliza una cámara Intel D435 [69] que es 3D y se sitúa en posición cenital. En el caso particular 2D la cámara utilizada es una cámara *comodity* usb de resolución 640 x 480 con 8bits de color.

Para realizar la cuantificación se propone un método que genera una magnitud en cada punto de las superficies a analizar. Dicha magnitud indica si este punto ha sido observado y en qué cantidad. Es decir, dicha magnitud es proporcional al tiempo de observación y a la distancia a la que se encuentra el punto observado por la persona.

De hecho, está relacionada con la zona del fondo de ojo que ha sido utilizada para observar el objeto y el tiempo que se ha observado. Por tanto, se utiliza la sensibilidad en función de la zona en la que se encuentra el objeto respecto al observador. Las trayectorias de las personas se calculan determinando la posición de la cabeza y la dirección de visionado. En este trabajo suponemos que la dirección natural del ojo coincide con la dirección de la cabeza.

En cada punto de la trayectoria se utiliza un modelo de la visión del ojo que transforma la zona de retina asociada a una magnitud en la superficie, tal como se verá en la descripción del método. Dicha magnitud la denominaremos densidad de foco y se calcula utilizando una nube de puntos PCL (Point Cloud Library) 3D [58] de las superficies a analizar, ya sean interiores o exteriores a la zona de captura de la trayectoria. La densidad de foco se acumula en cada frame de la trayectoria. Finalmente, se determina la atención en los objetos sumando

la densidad de foco en cada punto de la superficie. Los que tengan mayor densidad de foco serán los que han generado mayor interés. Dichas magnitudes, no tienen un valor absoluto sino un valor de comparación entre diferentes objetos de interés.

Se separa el método de búsqueda del foco de atención de la implementación práctica realizada, dado que pueden existir multitud de implementaciones, ya sean con cámaras cenitales o con cámaras frontales, con cámaras de bajo coste o de alta precisión, pero éstas no modifican la validez del método de determinación del foco de atención utilizado en este trabajo. El método de estimación de los parámetros puede cambiar mucho la precisión del resultado dependiendo del ruido de las cámaras, así como del número y de su disposición espacial.

Para verificar el método se utiliza una cámara con sensores inerciales (IMU: *Inertial Measurement Unit*) ubicada en la cara del individuo mediante un dispositivo. Para no confundir a esta cámara nos referimos como la cámara frontal. Con dicha cámara se determina que está viendo cada individuo en cada instante de la trayectoria. Además, con la IMU asociada a la cámara se determinan los ángulos yaw, pitch y roll de la cabeza.

Se dispone igualmente de unos pósteres de color colocados estratégicamente en mesas y paredes para realizar la verificación. Finalmente, con las imágenes obtenidas de la cámara frontal (la que simula el ojo), se calcula para cada póster la cantidad de pixeles que ha ocupado. Estas se suman en cada *frame,* y este resultado se compara con el resultado de la densidad de foco asociado a los diferentes pósteres. Como podemos ver, el método de validación utiliza la cámara como si fuese un ojo para confirmar el resultado.

3.2 ZONA DE ANÁLISIS

En la Figura 1 se muestra la configuración experimental con las dimensiones de la sala, la zona que capta la cámara cenital (donde hay imagen) y los objetos de interés. Los objetos de interés son las tres láminas con un marco de otro color situadas encima de la mesa. También lo forman los cuatro carteles: rojo, naranja, verde y blanco, situados en las paredes. Aunque los carteles están fuera del campo de visión de la cámara situada en el techo no es un problema para el algoritmo propuesto. Los carteles situados encima de las mesas complementan a los de la pared. La razón principal es que mientras el ángulo de visión de la

pared es próximo a 90º los carteles de las mesas tienen ángulos que van desde 160º a 100º y de esta forma las pruebas de validación disponen de una amplia gama de ángulos de visión.

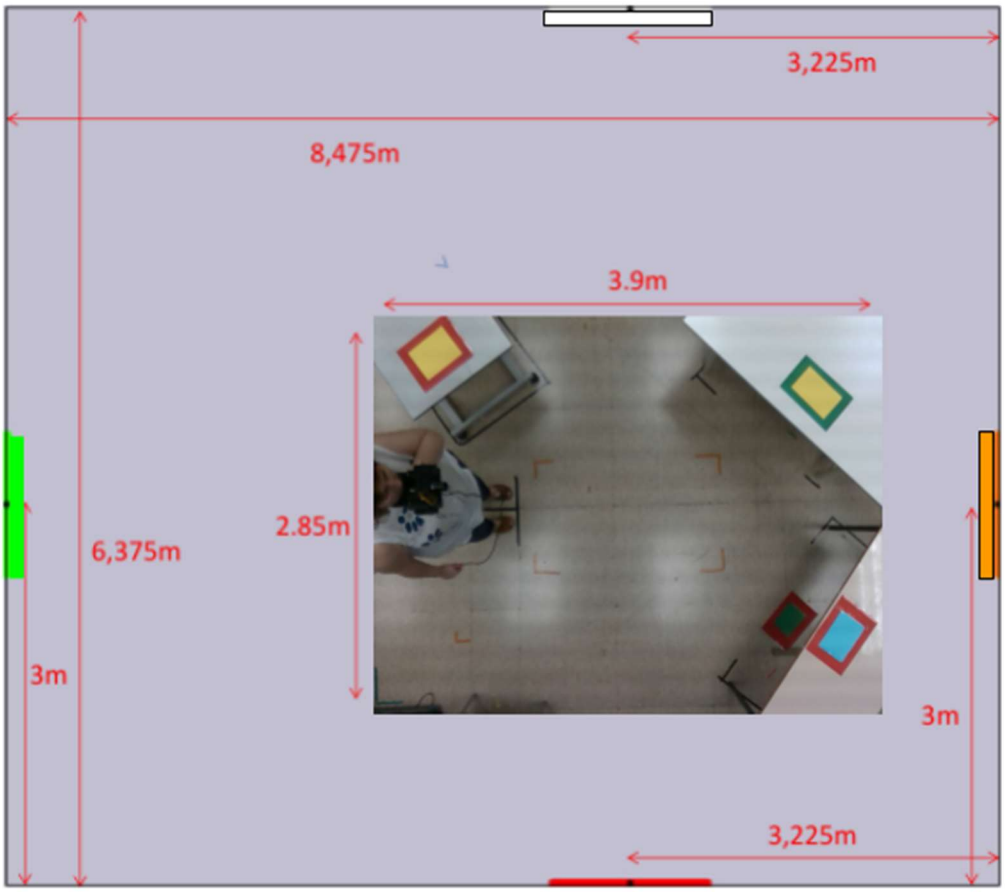


Figura 1. Configuración experimental

Toda la zona que se muestra en la Figura 1 se convierte a una nube de puntos donde se realizará el análisis (Figura 2).

Lo primero que se puede observar en la Figura 2 es que hay zonas con mayor densidad de puntos que otras. Esto se hace así para acelerar el tiempo de procesado, que depende del número de puntos. Sin embargo, aunque haya zonas con menor densidad de puntos se mantiene un detalle aceptable de la zona a analizar.

Obsérvese que la densidad de puntos es mayor en la zona de captura de la cámara (zona amarilla) y menor en la zona externa. Hay objetos de análisis situados en una zona bastante densa y otros, los de las paredes, en una zona con menor densidad de puntos.

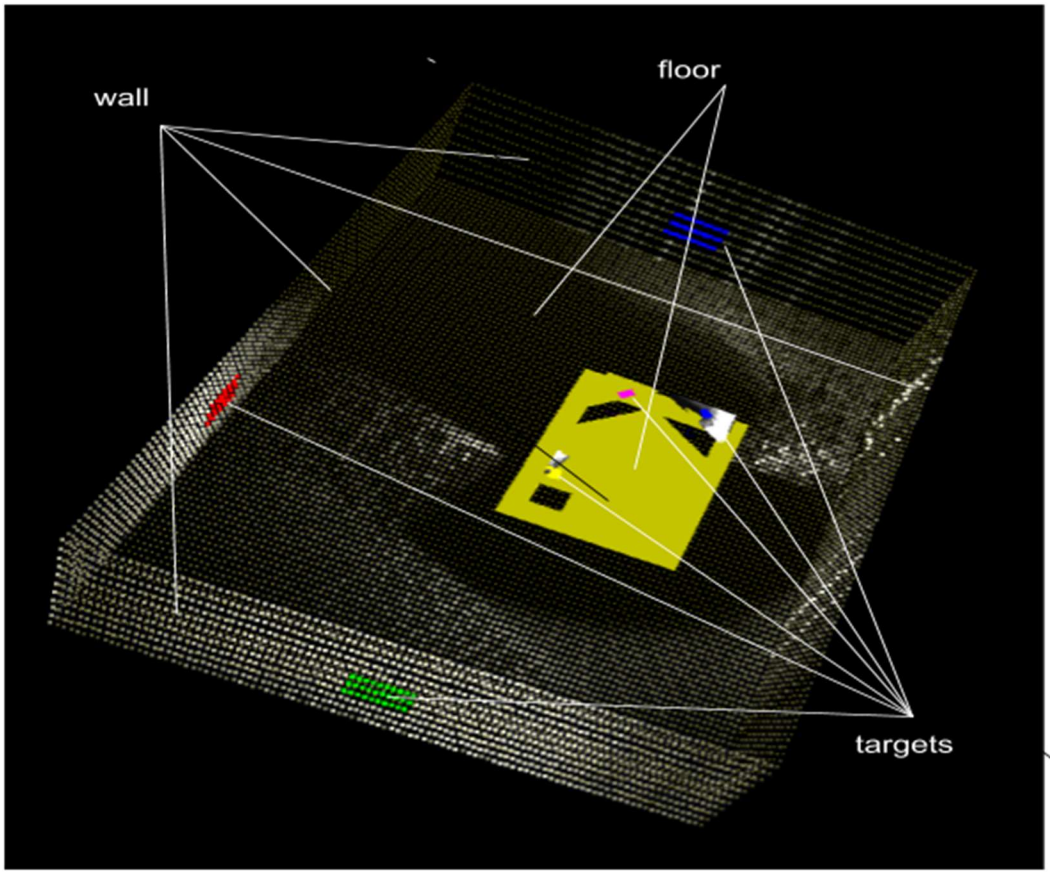


Figura 2. Distribución de nube de puntos de la zona de análisis. En amarillo, la zona capturada por la cámara, que dispone de alta densidad de puntos, mientras que las paredes y el suelo tienen una densidad reducida. Los objetivos de análisis están en la zona de cámara, y en las paredes, fuera de ella.

3.3 TRAYECTORIAS

Como primer elemento en nuestro procedimiento definiremos la *trayectoria* de las personas. Una *trayectoria* está formada por la secuencia temporal de la posición de la cabeza y los ángulos que determinan la dirección de visionado (Figura 4). Por tanto, dicha *trayectoria* se caracteriza por una sucesión ordenada de posiciones y ángulos, que localizan la cabeza en el espacio y la orientan respecto a su entorno:

$$T_r = \{P_n\} \tag{1}$$

donde

$$P_n = \{(x, y, z), (yaw, pitch, roll)\} \quad (2)$$

(x, y, z) es su posición y el segundo término son los tres ángulos: inclinación (*pitch*), giro vertical (*roll*) y giro horizontal (*yaw*) [43]. Para clarificarlos en la Figura 4 se muestran los tres grados de libertad de la cabeza humana que se corresponden con los tres ángulos.

En la Figura 3 se muestra un ejemplo de trayectoria donde la posición a lo largo del tiempo se marca en color blanco, y en color rojo están las flechas que indican los ángulos. La posición y las flechas se calculan para cada *frame* y su valor depende de la velocidad de captura de la cámara.

Figura 3. Ejemplo de trayectoria donde se indican las posiciones y ángulos de la cabeza para cada frame.

Para obtener el foco de atención en 2D se ha de tener en cuenta sólo el ángulo de giro horizontal, *yaw*. Sin embargo, si es 3D se ha de considerar los ángulos *yaw* y de inclinación *pitch*. Tenemos en cuenta sólo los ángulos *pitch* y *yaw* porque son los que nos interesan para poder determinar el foco de atención de la gente en un recinto. El objeto de atención puede estar a diferentes alturas, así como las personas pueden tener diferentes alturas, por tanto, si

analizamos el ángulo *pitch* estamos introduciendo otro grado de libertad que nos permitirá aumentar el rango de aplicaciones del sistema.

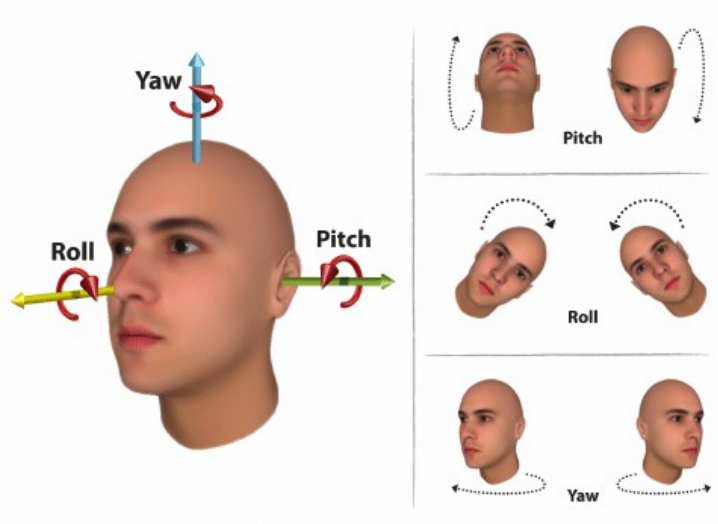


Figura 4[1]. Orientación de la cabeza en términos de movimientos de inclinación (pith), giro vertical (roll) y giro horizontal (yaw) que describen los tres grados de libertad de una cabeza humana

La Figura 3 es un ejemplo de trayectoria donde se indican las posiciones y ángulos de la cabeza para cada frame. De la Figura 3 se irían determinando los:

$$\begin{aligned} P_o &= (x_0, y_0, z_0), \{(yawn_0, pitch_0)\} \\ P_1 &= (x_1, y_1, z_1), \{(yawn_1, pitch_1)\} \\ &\vdots \\ P_n &= (x_n, y_n, z_n), \{(yawn_n, pitch_n)\} \end{aligned} \tag{3}$$

Ese conjunto de valores nos proporcionaría la trayectoria T_r.

$$T_r = \{P_0, P_1, \dots P_n\} \tag{4}$$

[1] Procedencia de la Figura: E. Arcoverde, R. Duarte, R. Barreto et al., " Enhanced real-time head pose estimation system for mobile device. Integrated Computer Aided Engineering" Vol.21, no 3. 281-293. 2014.

3.4 MÉTODO 3D

Para determinar la atención que una persona presta a un objeto vamos a analizar cómo trabaja el sistema visual humano, más concretamente los ojos.

El ojo humano dispone de un sensor, la retina (Figura 5), donde se forman las imágenes proyectadas a través de una lente, el cristalino (Figura 5). Para determinar cuál es la atención de un objeto debemos determinar qué porcentaje está ocupando dicho objeto en la retina y durante cuánto tiempo. La zona ocupada en la retina la mediremos en ángulo sólido (estereorradianes) relativo al foco. Esta zona ocupada en la retina dependerá de la superficie del objeto en el mundo real y de su distancia al ojo (Figura 6). La Figura 6 ilustra que un objeto observado que está a una distancia superior a dos veces la distancia focal, la imagen que se forma está invertida y es de tamaño menor al tamaño real. Así mismo, objetos de dimensión distinta producirán la misma atención (misma zona de excitación del ojo) si estos objetos están a distancias distintas del ojo, pero son las adecuadas.

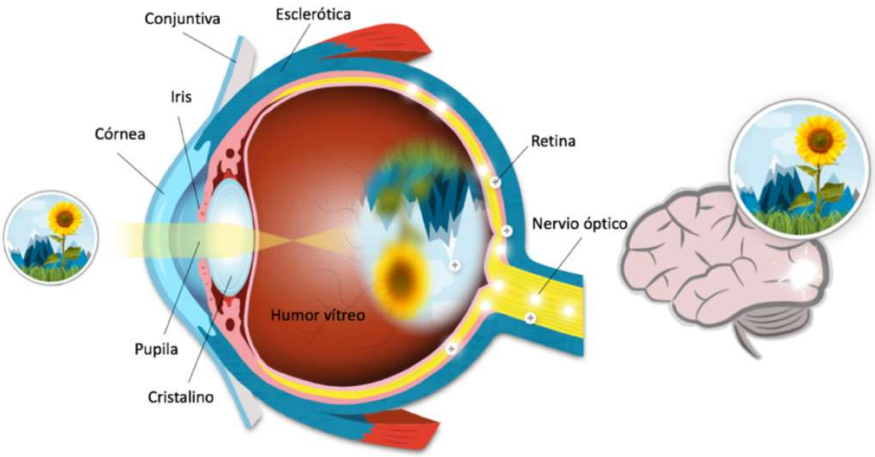


Figura 5. Partes del ojo. Imagen adaptada a partir de: https://www.visiondirect.es/ojo-humano[2].

[2] "Problemas de visión: consejos y recomendaciones | #desprésdelcàncer." [Online]. Available: https://despresdelcancer.cat/index.php/problemes-visio-consells-recomanacions/?lang=es. [Accessed: 07-Jul-2020]

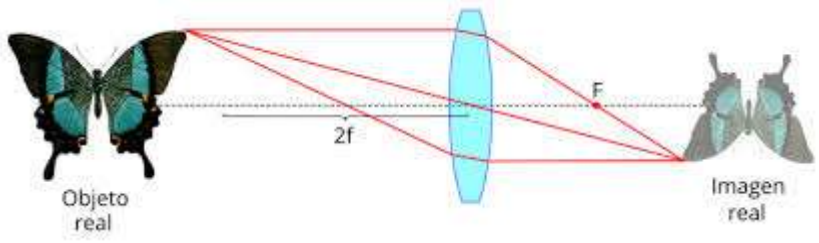


Figura 6. Imagen formada a través de una lente convergente de distancia focal f.

La relación de ángulo sólido constante viene dada por:

$$S_r = \frac{A}{d^2} \tag{5}$$

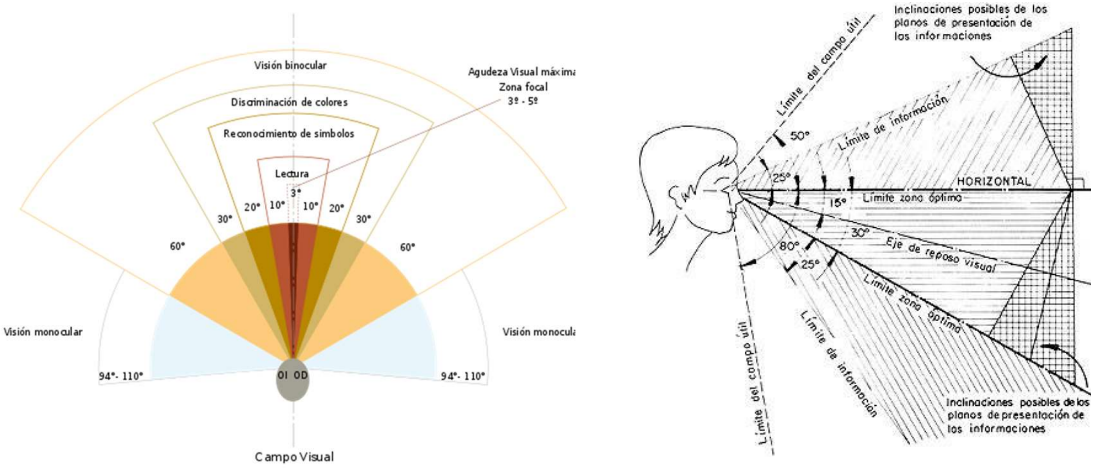
donde A es el área del objeto y d la distancia al ojo. Por tanto, para mantener la misma atención en el ojo el tamaño del objeto debe crecer con el cuadrado de la distancia. Además, la atención humana tiene dependencia angular horizontal (Figura 7 b)) y vertical (Figura 7 a)).

Esta dependencia angular la indicaremos mediante una función multiplicativa dependiente del ángulo. En el caso del ojo el foco de atención[3] se puede definir,

$$F_{ángulo}(\alpha, \theta) = Ge^{\left(-\left(\frac{\alpha_x^2}{2\sigma_x^2}+\frac{\theta_y^2}{2\sigma_y^2}\right)\right)} \tag{6}$$

Donde α y θ son los ángulos con el eje x e y, respectivamente, en el sistema de referencia del ojo, σ_x y σ_y están asociados a los ángulos del campo de visión (horizontal y vertical), así, por ejemplo, si nuestro interés es analizar cuál ha sido la atención con capacidad de lectura, el ángulo seria de ±10º y σ=10.

[3]La sensibilidad del ojo va decreciendo con el ángulo tanto vertical como horizontal (Figura 7) por lo que la expresión anterior es una modelización de la sensibilidad encontrada con medidas de campo visual.



a) b)

Figura 7. Campo de visión dependiendo del ángulo[4, 5]: a) horizontal y b) vertical[6].

Si en lugar del ojo utilizásemos una cámara,

$$F_{ángulo}(\alpha,\theta)=\begin{cases}0, & si\ \ ángulo > ángulomáximo\\ 1, & enotroscasos\end{cases} \tag{7}$$

En las cámaras los ángulos dependen de que el formato sea 16:9 o 4:3. Este punto es pertinente en este caso debido a que la verificación del método se realizará con una cámara frontal, en la que no podremos aplicar la ecuación correspondiente al ojo humano. Principalmente porque en el ojo del que hemos comentado su modelo anteriormente las células sensoriales no tienen una distribución uniforme, hay más en el centro y menos en la periferia; mientras que en la cámara las células sensoriales tienen una distribución uniforme y rectangular.

Al objeto de clarificar la expresión (6) mostramos la Figura 8 donde se puede observar desde el sistema de referencia del ojo como serían las diferentes variables.

[4]"Campo visual" [Online]. Available: https://www.ecured.cu/Campo_visual [Accessed: 23-Jul-2020].

[5] "Estudio de los índices del campo visual" [Online]. Available: https://www.tdx.cat/bitstream/handle/10803/4249/ags02de13.pdf [Accessed:23-Jul-2020]

[6] "El campo de visión." [Online]. Available: http://www.racesimonline.com/articulos/El_campo_de_vision.php. [Accessed: 07-Jul-2020].

Observador localizado en el punto $P_i=\{r_i,a_i\}$ donde:

r_i: punto (x_i,y_i,z_i) donde se haya el observador

a_i: vector de dirección de la mirada del observador según la posición de cabeza y ojos.

Para evaluar la densidad de foco de atención del observador al objeto hay que dividir el objeto en pixeles o elementos mínimos de superficie del objeto orientado al observador. En la Figura 8 se puede ver que la distancia del observador al pixel considerado vale *d* (es el módulo del vector r_d). Proyectando el objeto y el pixel considerado en el plano ZX se puede ver la proyección del vector r_d, r_{dx} y el ángulo θ_x que representa el ángulo horizontal de la Figura 7 b) que se usa en la función F_{angulo} (6). Al proyectar sobre el plano ZY se ve el vector r_{dy} y el ángulo θ_y que representa el ángulo vertical de la Figura 7 a) y también de la función F_{angulo}.

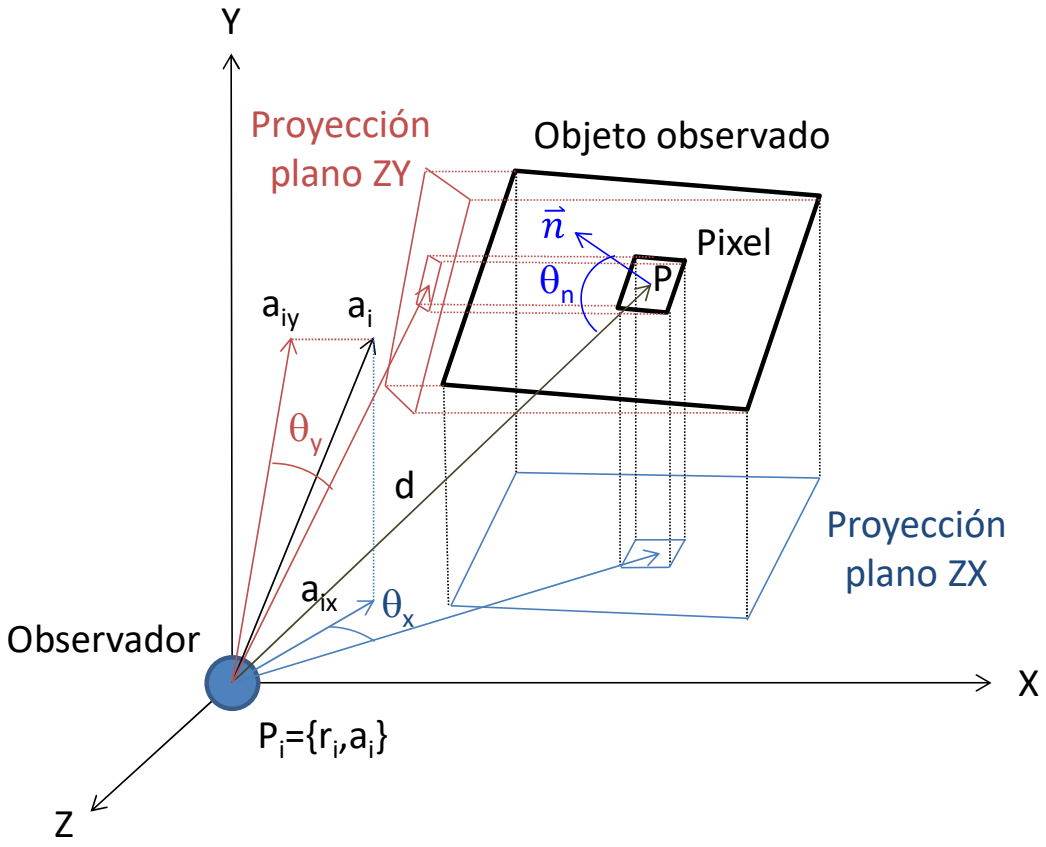


Figura 8. Esquema de un Observador mirando un objeto.

Como se ha comentado anteriormente el foco de atención es proporcional al ángulo sólido del objeto proyectado en la retina del ojo. Por tanto, una parte de la función densidad de atención será una función proporcional a dicho ángulo. Dado que la F_{angulo} está relacionada con la superficie del objeto proyectada en la retina, la densidad de foco será proporcional a F_{angulo} e inversamente proporcional al cuadrado de la distancia d, tal como indica la ecuación del ángulo sólido (5). Por último, hay que considerar que, si la visión no es frontal, la información de la atención se debe penalizar con la función coseno del ángulo normal a la superficie, es decir, θ_n. Con todas estas consideraciones la función de densidad de atención *DFOA* (DFOA: *density focus of attention*) en un punto determinado P la definimos de la siguiente manera:

$$DFOA(P) = K \cdot F_{\acute{a}ngulo} \cdot \frac{1}{d^2} \cdot cos\theta_N \tag{8}$$

donde K es la constante de normalización, $F_{ángulo}$ la función descrita en la ecuación (6) o (7) , d la distancia desde el ojo al punto P y θ_N el ángulo con la normal a la superficie donde reside el punto P. Si bien se puede complementar los factores con funciones que premien la dirección de movimiento, así como la posición relativa de la cabeza, en posiciones naturales o forzadas, desgraciadamente no son medibles de forma objetiva; por lo que no las utilizaremos en este trabajo.

Se puede calcular la *DFOA(P)* en cualquier punto P perteneciente a una superficie (suelo, paredes, etc…) o a un objeto presente en la zona de análisis. Pero no tiene sentido la *DFOA(P)* en puntos del aire. La función *DFOA(P)* valdrá cero en puntos de objetos ocultos para el observador y en los puntos que el observador no ha mirado durante el tiempo de análisis.

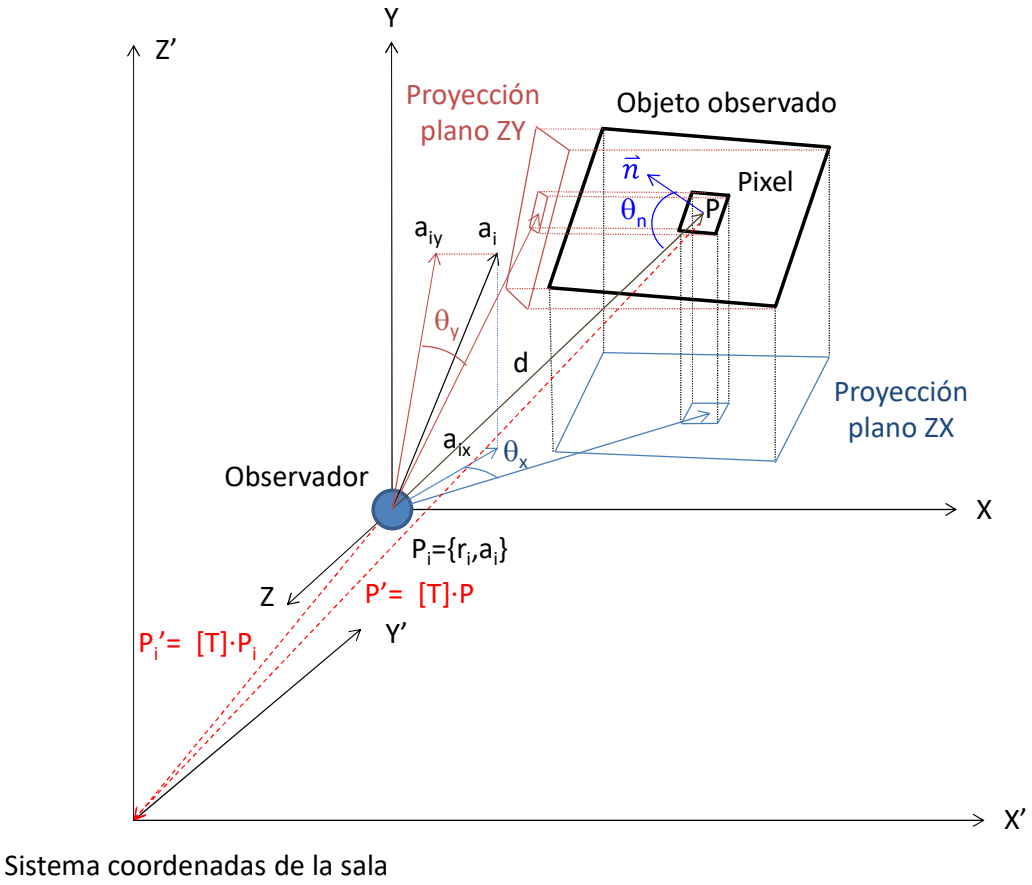


Figura 9. Sistema de coordenadas de la sala respecto al sistema de referencia del ojo.

La *DFOA(P)* calculada hasta ahora está basada en el sistema de referencia del ojo, y puesto que la posición y ángulo del ojo está cambiando en cada posición de la trayectoria es muy complicado de manejar. Para solventar este punto realizaremos una transformación ([T]) al sistema de referencia de la sala (X'Y'Z') (Figura 9),

$$DFOA(P') = [T] \cdot DFOA_{local}(P) \tag{9}$$

donde $DFOA_{local}(P)$ es la *DFOA* calculada en el sistema de referencia del ojo del observador, y *DFOA(P')* es la *DFOA* transformando el punto P del sistema de referencia del ojo al punto P' del sistema de referencia de la sala.

Sin embargo, la *DFOA(P')* solo indica el foco de atención de un individuo desde un punto determinado. Para calcular el foco de atención visual en un punto (*VFOA*: *Visual Focus of*

Attention) se tiene que realizar la suma de todas las *DFOA(P')* en todos los puntos de la trayectoria (T_r) del individuo. Además, se debe añadir una suma de todos los individuos de un colectivo (C) considerados en el análisis. Para normalizar la *VFOA(P')*, se multiplica las sumas por una constante N:

$$VFOA(P') = N \sum_{C} \sum_{T_r} DFOA(P') \tag{10}$$

La normalización se realiza calculando la suma de todas las *VFOA(P')* de todos los puntos *P'* de los objetos y superficies de la zona de análisis, y ajustando N para que esa suma valga 1:

$$1 = \sum_{P} VFOA(P') = N \sum_{P} \sum_{C} \sum_{T_r} DFOA(P') \tag{11}$$

$$N = \frac{\sum_P VFOA(P')}{\sum_P \sum_C \sum_{T_r} DFOA(P')} \tag{12}$$

La expresión *VFOA* es una nube de puntos de intensidad que, por ejemplo, se puede representar mediante un código de colores. Se ha escogido el negro para representar máxima atención y azul para indicar poca atención. En la Figura 10 puede verse como la persona en su trayectoria presta máxima atención a las mesas de la zona de captación que son las que están de color negro y al resto poca atención.

La ecuación (10) no incluye el tiempo, sin embargo, se tendría que tener en cuenta cuando consideramos múltiples trayectorias simultáneas. Si tenemos en cuenta la variable tiempo la ecuación (10) quedaría:

$$VFOA(P', t) = N \sum_{C} \sum_{T_r(t)} DFOA(P', t) \tag{13}$$

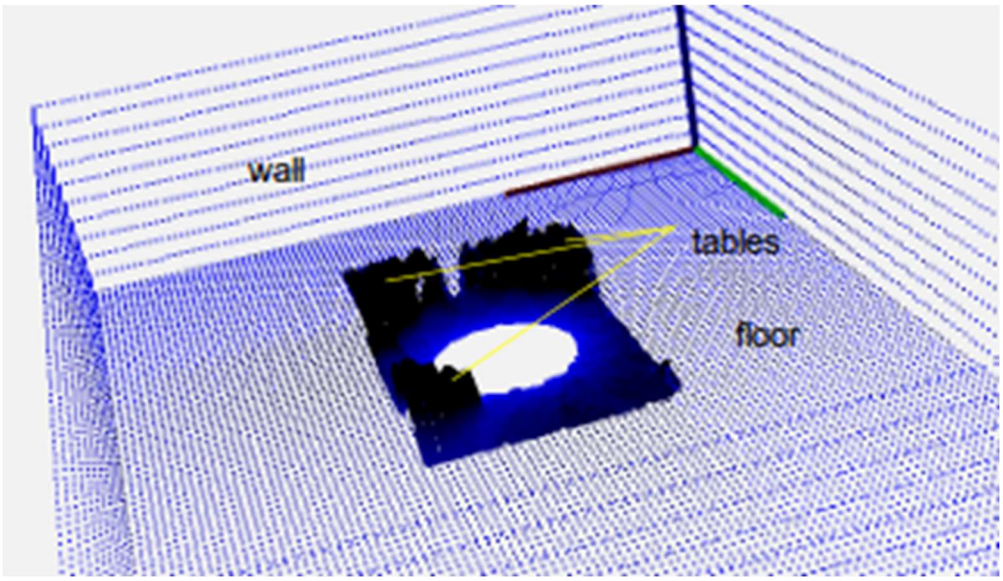


Figura 10. Distribución de la atención humana en una trayectoria

La $VFOA(P',t)$ expresa la dependencia temporal de la *VFOA*, donde los sumatorios se realizan en un colectivo determinado de las *DFOA* que coincidan con el tiempo t. Naturalmente, esto solo tiene aplicaciones no triviales en el caso de que existan múltiples trayectorias simultáneas. En estas trayectorias las sumas se realizan dentro de todos los elementos que en el instante de tiempo t hayan estado en la zona de tránsito.

La expresión de la *VFOA* acumulada, puede también restringirse a un intervalo, entre t1 y t2:

$$VFOA\big(P, t(t1, t2)\big) = N \sum_{t1,t2} VFOA(P,t) \tag{14}$$

La principal función de esta dependencia temporal es la sincronización de la función de foco de atención con eventos ocurridos en la zona de análisis, de forma que se pueda determinar qué evento ha producido más atención.

El método global, sin dependencia temporal es sobre el que hemos realizado las comprobaciones de validación.

3.5 MÉTODO 2D

El método 2D es una simplificación del método 3D que se puede utilizar en implementaciones de menor coste de hardware y computacional.

Se basa en utilizar solo la información contenida en la proyección de los objetos de la sala y de la trayectoria de los observadores sobre el plano X'Y' del sistema de coordenadas de la sala que se muestra en la Figura 9. Por tanto, en lugar de la posición (x',y',z') y los tres ángulos de Euler (yaw, pitch, roll), se usa la posición (x',y') y un solo ángulo de Euler (yaw).

Para el cálculo de la *DFOA(P)* en el sistema de coordenadas del ojo se usa solo la proyección sobre el plano ZX del sistema de coordenadas del observador que se muestra en la Figura 8. Posteriormente se realiza el cambio de variables al sistema X'Y' de la sala.

Los cálculos con estas trayectorias 2D son más fáciles de conseguir, aunque también tendrá limitaciones en las informaciones que se podrán extraer y en el ámbito de utilización.

En este método las funciones tienen un ligero cambio en su estructura. Partimos también del funcionamiento del ojo humano, pero en este caso solo tenemos en cuenta el campo de visión horizontal y consideramos el valor medio del campo de visión vertical. En la Figura 11 se puede observar la integración sobre el eje vertical de la imagen de la manzana, se observa que solo ciertas líneas verticales tienen información mientras el resto no.

En la gestión del foco de atención 2D se considera como si la retina fuese un sensor lineal horizontal. Por tanto, el efecto que producirá en el ojo cualquier estímulo dependerá del ángulo horizontal que ocupe el mismo, en relación con el campo de visión. Es decir, en lugar de un ángulo sólido, como manejamos en el método 3D (5), ahora trataremos con un ángulo lineal A:

$$A = \frac{L}{d} \tag{15}$$

donde L es la anchura horizontal del objeto y d la distancia del ojo al mismo.

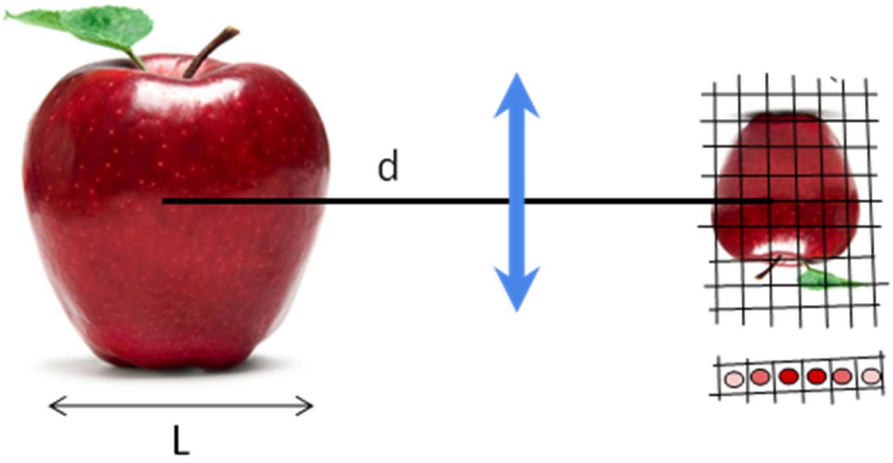


Figura 11 .Campo de visión vertical

La función de densidad de foco de atención (*DFOA*) para 2D en un punto determinado *P* sería la siguiente:

$$DFOA_{local}(P) = K \cdot F_{ángulo} \cdot \frac{1}{d} \cdot cos\theta_N \tag{16}$$

donde K es la constante de normalización, d la distancia desde el ojo al punto P y θ_N es el ángulo con la normal a la línea donde reside el punto.

Y la $F_{ángulo}$ tendría la siguiente expresión:

$$F_{ángulo}(\alpha) = Ge^{\left(-\left(\frac{\alpha_x^2}{2\sigma_x^2}\right)\right)} \tag{17}$$

Y la siguiente para el caso de una cámara:

$$F_{ángulo}(\alpha) = \begin{cases} 0, & ángulo > el\ ángulo\ máximo \\ 1, & en\ otro\ caso \end{cases} \tag{18}$$

Igual que en el caso anterior se hace la transformación [T] al sistema de referencia de la sala,

$$DFOA(P') = [T] \cdot DFOA_{local}(P) \tag{19}$$

Las funciones *VFOA* mantienen la misma estructura que en el caso 3D, pero se transforma a unos ejes de coordenadas X'Y' que son la proyección al suelo horizontal de la sala, es decir, la trayectoria del observador es T_r'.

$$VFOA(P') = N \sum_{C} \sum_{T_r'} DFOA(P') \tag{20}$$

La Figura 12 representa, en escala de color, la magnitud *VFOA* generada mediante el análisis de varias trayectorias.

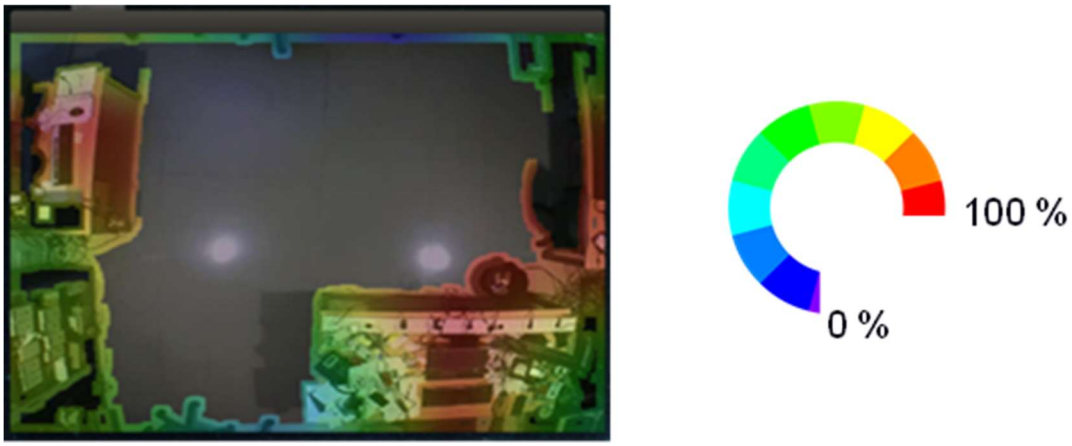


Figura 12. VFOA identificada en escala de color. Si está en rojo es máxima la VFOA y si está en azul oscuro es que esa zona no interesa nada

3.6 TIEMPOS DE ATENCIÓN

Para el análisis de los tiempos de visualización consideramos tres definiciones de tiempo de las trayectorias orientadas:

- Tiempo de permanencia (*Dwell time)*
- Tiempo de visualización (*In-View time*)
- Tiempo de atención (*Attention Time*)

Las trayectorias orientadas se calculan determinando en cada cuadro o frame grabado desde una cámara en posición cenital, la posición (x', y') y la orientación de la cabeza del usuario en relación con las coordenadas de la sala. Una persona dentro de la sala se puede parametrizar usando un vector de estado X:

$$X = [p', \psi, v, \phi] \tag{21}$$

donde:

p' = (x',y') indica la posición de la persona en el sistema de coordenadas de la sala.

ψ es la dirección definida por la trayectoria de la persona.

v es la velocidad instantánea de la persona.

ϕ es el ángulo de la cabeza, también en el sistema de coordenadas de la habitación (ver Figura 13 a)).

Una trayectoria orientada T se define como la secuencia temporal de estados para todos los instantes k en los que una persona está en el campo de visión de la cámara:

$$T = \{X_k\} \tag{22}$$

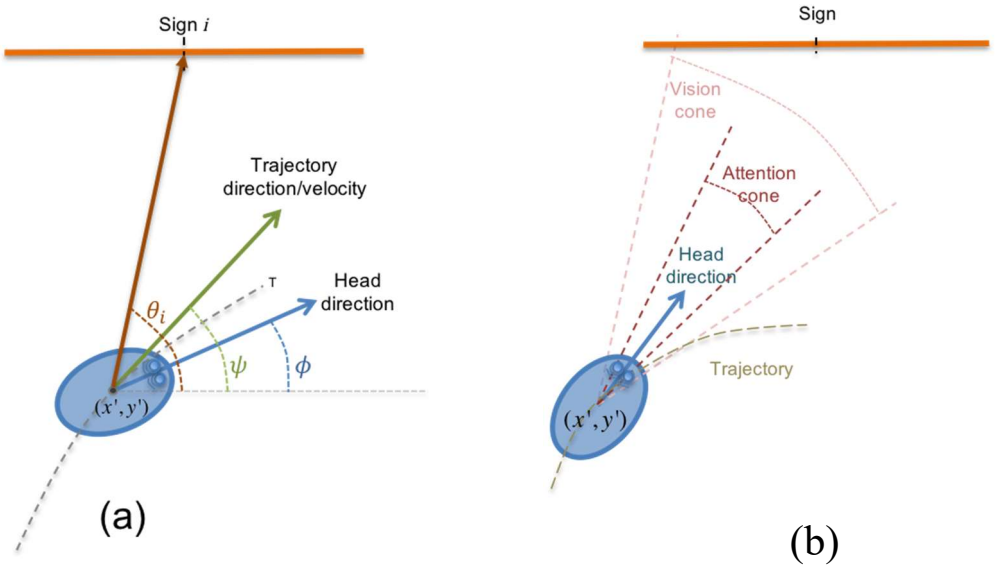


Figura 13. a) Variable de estado b) Ángulo de visión de In-View y de atención

La determinación de las métricas temporales atención (*Attention),* tiempo de visualización *(In-View) y* tiempo de permanencia *(Dwell-time)* viendo el objeto se basan únicamente en los ángulos θ_k, ϕ_k obtenidos en cada instante de tiempo.

El **tiempo de permanencia (*Dwell time*)** es la cantidad total de tiempo que la persona es visible por la cámara colocada en el punto de interés; que cuando se visualiza desde la cámara cenital equivale a un ángulo de la cara con respecto al cono de visión (Figura 13 b)).

El **tiempo de visualización (*In-View time*)** se puede definir como el tiempo durante el cual el usuario puede ver el objeto de interés con un ángulo de cono de atención.

Hay varias definiciones del tiempo de visualización. Por ejemplo, la guía de *Media Rating Council* (MRC) [59] considera que el tiempo de visualización es cuando más del 50% del objeto de interés está a la vista durante más de 1 segundo. Por otro lado, [50] define el tiempo de visualización como el tiempo que una cámara situada en lo alto del objeto de interés puede detectar la cara del cliente usando un detector frontal del rostro.

Utilizaremos un enfoque similar a [28], ya que se puede relacionar con los ángulos medidos θ, ϕ. De hecho, un típico detector de cara frontal como el one de OpenCV [70][71] puede detectar caras no frontales, hasta ángulos aproximadamente de 45º en ambas direcciones. Esto es equivalente a considerar que un cliente tiene un anuncio en *In-View* cuando se cumpla $|\theta - \phi| \leq 45º$ (Ver la Figura 13). Para evitar detecciones falsas mantendremos el requisito establecido en [72] de que el cartel debe estar a la vista durante más de un segundo. Para esto, analizaremos la secuencia temporal $\{X_k\}$, clasificando cada instante como *In-view* (X_k^{iv}) o *no In-view*. Extraemos las sub-secuencias en las que todos los ángulos consecutivos se clasifican como *In-view* y que son más grandes que un segundo $\{X_k^{iv}\}_j$ (j es el índice de la sub-secuencia). Luego, para una trayectoria dada i, el tiempo de visualización es:

$$T_i^{iv} = \sum_j longitud\left(\{X_k^{iv}\}_{i,j}\right) \cdot t_f \tag{23}$$

donde t_f es la duración del instante que depende sólo del bitrate del video.

Para el cálculo del **tiempo de atención (*Attention time*)**, el requisito es que el cliente esté activo mirando el anuncio. En [2], esto se determinó utilizando un modelo activo de

apariencia (AAM) de la cara para estimar la dirección de la mirada del cliente. Nuestro enfoque equivalente es considerar los casos donde |θ - ϕ| son lo suficientemente pequeños para que el anuncio se encuentre dentro del cono de visión donde el cliente es capaz de tener toda la atención, esto es, cuando |θ - ϕ| ≤ 25º. Se ha derivado este valor a partir de estudios de fisiología del campo visual [62]. Como hemos hecho anteriormente, clasificamos cada instante como atención o no atención y mantenemos las subsecuencias de instantes de atención consecutivos que tienen más de un segundo. El tiempo de atención es la suma sobre todas las trayectorias i:

$$T_i^a = \sum_j longitud\left(\{X_k^a\}_{i,j}\right) \cdot t_f \tag{24}$$

Al sumar todas las trayectorias, se pueden calcular los valores finales de los tiempos de atención (T^a) y de visualización (T^{iv}):

$$T^a = \sum_i T_i^a \ , \qquad T^{iv} = \sum_i T_i^{iv} \tag{25}$$

3.7 MÉTODOS 3D, UTILIDADES Y LIMITACIONES

3.7.1 UTILIDADES

Existen multitud de aplicaciones del método propuesto; todas donde queramos realizar comparaciones objetivas de la atención prestada por un colectivo o por individuos, algunos ejemplos podrían ser los siguientes:

1. COMPARACIÓN OBJETIVA DE OBJETOS

La principal aplicación de los métodos de detección de foco de atención está en la comparación de la atención suscitada por diferentes objetos en la zona de análisis. De hecho, el método permite comparar de forma objetiva las predilecciones de ciertos objetos con respecto a otros de un determinado colectivo.

Esto permite establecer campañas de publicidad con un objetivo mucho más dirigido o así mismo determinar que objetos van a tener mayor aceptación desde el punto de vista comercial. También podría extrapolarse a la valoración de obras de arte en base al interés mostrado.

2. VALORACIÓN OBJETIVA DE LOCALIZACIONES

Permite también determinar qué localizaciones son las más adecuadas para fomentar la venta de productos, utilizando el foco de interés. Así esas localizaciones son susceptibles de cobrar un coste mayor para colocar los productos.

3. AUTOMATIZACIÓN EN BASE A INTENCIÓN

El método presentado permite informar a otros equipos del interés suscitado y realizar acciones automáticamente, como llamar a una puerta o cambiar la luz de un semáforo, todas ellas de forma automática. Así por ejemplo la apertura de puerta automática de unos grandes almacenes requeriría no solo la presencia sino la intención de acceder, minimizando de esta forma las falsas aperturas que se realizan en los sistemas actuales.

4. TARIFICACIÓN DE LA PUBLICIDAD POR OBSERVACIÓN

La publicidad, que en muchos casos solo se tarifica por localización, podría modificarse en base al número de personas que han visualizado el anuncio. Esto ya que ocurre en el mundo de internet y en las pantallas de publicidad interactiva, pero no tanto en el mundo de la cartelería fija. Al mismo tiempo esto se podría realizar minimizando el número de cámaras.

5. VALORACIÓN OBJETIVA DE LAS PRESENTACIONES

Otra utilidad del método podría ser la valoración del interés de una presentación o clase sobre un colectivo; observando el seguimiento de ésta; o viceversa qué sujetos del colectivo están prestando atención.

3.7.2 VENTAJAS Y LIMITACIONES

El método 3D no tiene limitaciones en sí mismo; si bien es necesario disponer de una zona de análisis (nube de puntos) suficientemente precisa. De hecho, cualquier método que quiera medir la atención requiere corregir las posiciones reales distorsionadas por las capturas de las cámaras. Así mismo requiere determinar las alturas reales de los objetos a analizar y las oclusiones producidas. Por tanto, el método presentado 3D que utiliza la imagen y su profundidad es el único método para determinar la atención de una forma precisa.

Es importante destacar que en la implementación sería aconsejable utilizar más de una cámara 3D con el objeto de no tener zonas de sombra que dificultan la detección de los ángulos y que complican el tracking de los individuos.

En resumen, el método 3D es el único que permite corregir las posiciones de trayectorias y tratar las oclusiones; por tanto, es necesario para poder tener medidas precisas en distancias cortas y en ambientes donde las oclusiones pueden ser cambiantes.

3.8 MÉTODOS 2D, UTILIDADES Y LIMITACIONES

3.8.1 UTILIDADES, VENTAJAS Y LIMITACIONES

Con relación en el método 2D las utilidades pasan por tener un método económico, pero aproximado para poder tener una idea de las zonas de interés de los objetos.

Las utilidades planteadas en el método 3D valen también; pero en este caso no se puede determinar ni zonas de oclusión ni se puede corregir las trayectorias; con lo que las medidas obtenidas se deben interpretar posteriormente.

En resumen, los métodos 2D se pueden tomar como una medida aproximada del foco de atención que es suficiente para discriminar el interés en zonas amplias y bien separadas.

Las instalaciones necesarias para el procesado 2D pueden realizarse con cámaras simples y trabajando directamente con procesadores de bajo coste. Con ello, el mercado potencial de estas aplicaciones es alto; aunque solo se podrá utilizar cuando solo se requiera comparaciones simples (como, por ejemplo, saber solo si el cliente ha mirado a derecha o izquierda).

4 IMPLEMENTACIÓN

4.1 INTRODUCCIÓN

Entendemos por implementación el conjunto de técnicas que permite determinar la trayectoria de un individuo, esto significa, en el caso 3D determinar la posición de la cabeza (x,y,z), y los ángulos yaw, pitch y roll. En el caso 2D la posición de la cabeza (x,y) y el ángulo yaw.

Es este capítulo, presentaremos, en primer lugar, la implementación que utiliza una cámara cenital que da una imagen de profundidad (*Depth)* (implementación 3D) y en segundo lugar las técnicas utilizando solo una cámara cenital de imagen (implementación 2D).

La razón por lo que la técnica 3D no elimina completamente la técnica 2D y se incluye en este trabajo es exclusivamente económica, como ya se ha mencionado hay algunas aplicaciones donde no se requiere precisión y no existen problemas de oclusión y en ese caso una técnica 2D puede extraer información aceptable.

Por otra parte, a diferencia del método que consideramos estable y bien fundado, estamos trabajando en versiones futuras de la implementación, que permitirán una mayor precisión.

4.2 IMPLEMENTACIÓN 3D

El procedimiento para el cálculo de la VFOA en 3D sigue los siguientes pasos:

1. Determinación de la posición de la cabeza
2. Determinación de los ángulos
   - Yaw
   - Pitch
   - Roll
3. Cálculo DFOA y VFOA
   - Cálculo de trayectoria

- Zona de análisis
- Técnica de nubes de puntos para cálculo de DFOA y VFOA

En [93] se ha publicado algunos de los pasos que se mencionan aquí. A continuación, vamos a desarrollar cada uno de estos pasos.

4.2.1 DETERMINACIÓN DE LA POSICIÓN DE LA CABEZA

Para determinar la posición de la cabeza hay que localizar el cuerpo y definir lo que es la cabeza y su localización. Se seguirán los siguientes pasos:

I. Imagen de profundidad para eliminar el fondo inicial sin personas: Máscara MFA

Partimos de la imagen de profundidad, donde se elimina el fondo inicial captado sin personas (Figura 14), eliminando las zonas de mobiliario fijas en la zona de captación de la cámara. Definimos la máscara MFA, *Mask Free area*

$$MFA = \begin{cases} 1 & es \quad suelo \\ 0 & es \quad mesa \end{cases} \tag{26}$$

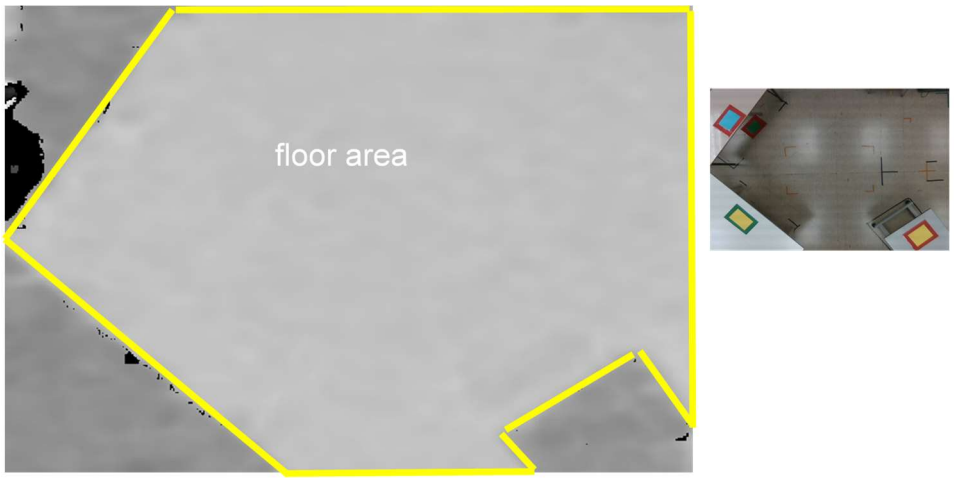
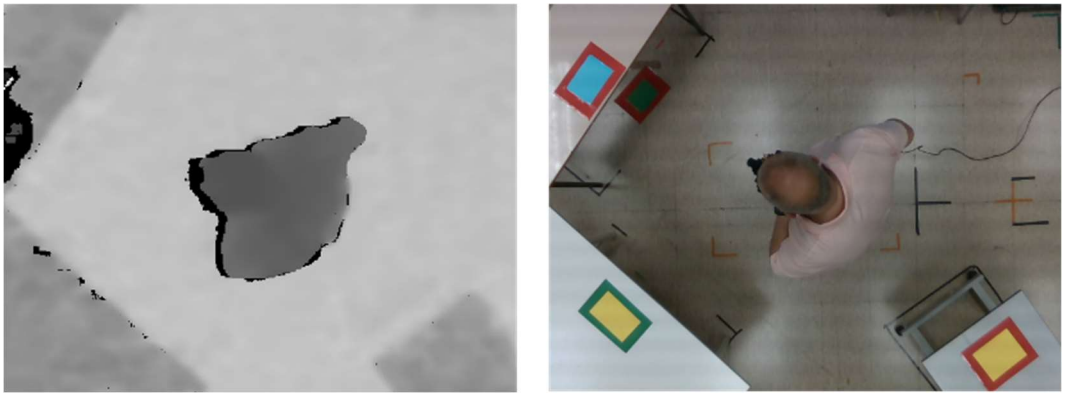


Figura 14. Imagen inicial sin personas que se utiliza como máscara de filtrado

En la Figura 15 se muestra a la derecha una imagen cenital RGB y a la izquierda de profundidad, en este caso realizando una trayectoria una persona. De hecho, suponemos que las personas solo pueden moverse por la zona donde hay suelo, el método no sería válido para movimientos que se pudieran producir por encima del mobiliario.



a) b)

Figura 15. A la derecha imagen cenital y a la izquierda imagen de profundidad

II. Histograma de profundidad: cálculo del umbral TRH

Para detectar las distintas partes de la imagen, Figura 15 (o similares), utilizamos el *histograma de profundidad* (Figura 16). En la Figura 16 se muestra el histograma donde se indica en qué lugar están situados cada uno de los elementos: suelo, persona y mesas. Esta distribución suele ser así para todas las imágenes captadas con la cámara en esa posición (Figura 15 b)).

Por tanto, sabiendo el tipo de histograma que se suele obtener podemos situar un umbral (TRH: *threshold*) para aislar a la persona como se puede ver en la Figura 16. Para determinar el punto de corte de final del cuerpo (TRH) hacemos el siguiente cálculo:

$$TRH = min(mayorquesuelo)(depth \cdot MFA) \quad (27)$$

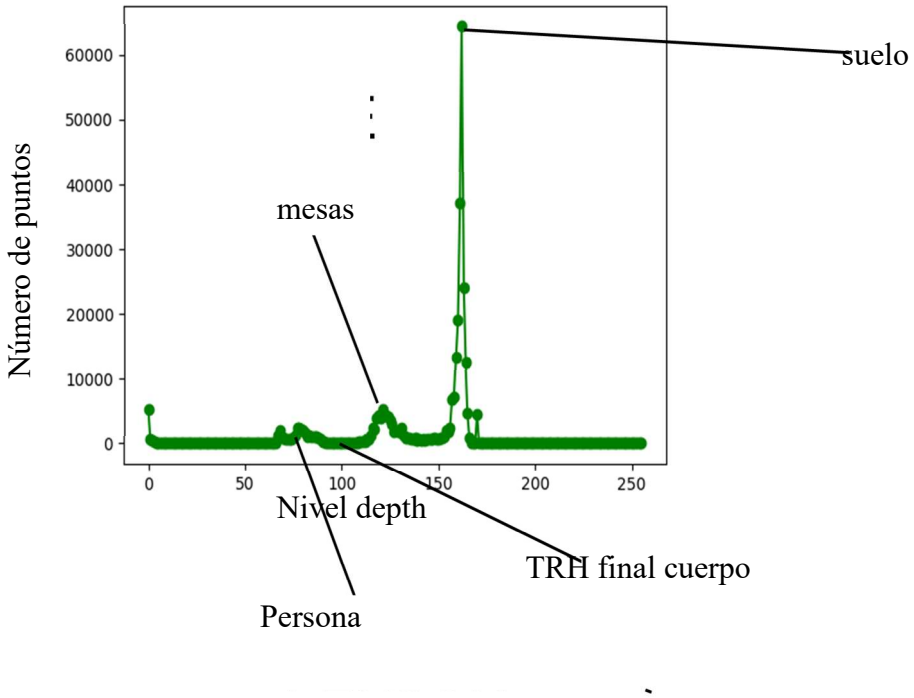


Figura 16. Histograma de la imagen de profundidad

TRH se determina en los primeros frames en donde se detecta a la persona para no hacer el cálculo largo y se mantiene durante el resto de la trayectoria.

III. Máscara de persona

A continuación, utilizando TRH se enmascara la imagen *depth* fijando a valor suelo todo lo que esté por debajo de TRH y lo que no se deja al valor que está,

$$Persondepth(x,y) = \begin{cases} depth.Inicial(x,y) & depth.Inicial(x,y) > TRH \\ Suelo & en\ otros\ casos \end{cases} \quad (28)$$

Aplicando la ecuación (28) se obtiene la imagen de la Figura 17.

IV. Validación por contornos

El siguiente paso es validar que esta imagen tiene una persona, para lo cual después de aplicar esta máscara se trazan todos los contornos, de todas las zonas que superen TRH como se aprecia en la Figura 18.

Dicha figura es un ejemplo de contornos que pueden aparecer después de comparar la imagen de profundidad con el umbral, TRH. Una parte se filtra usando el *background* y otra usando la superficie (Figura 18).

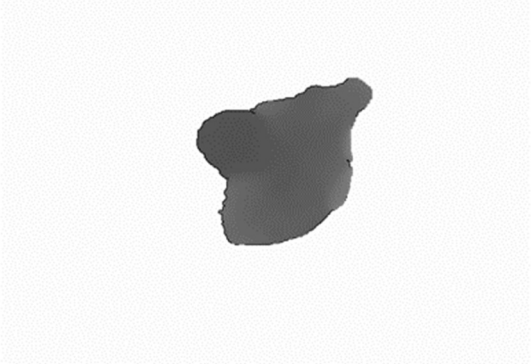


Figura 17. Imagen filtrada

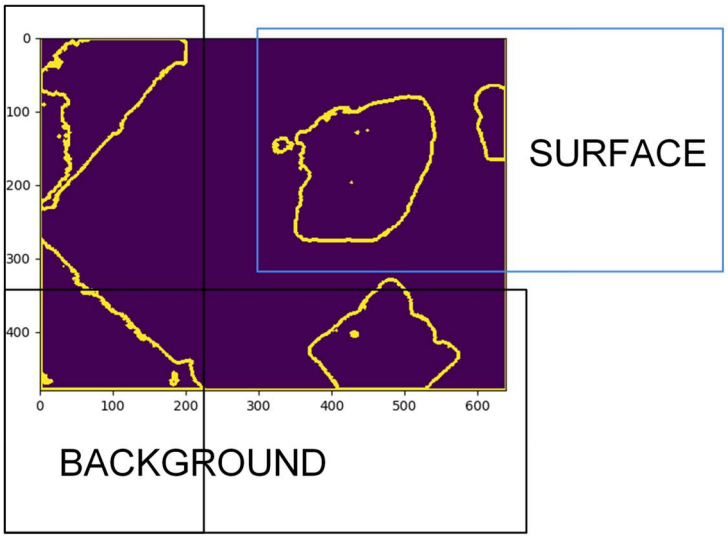


Figura 18. Ejemplo de contorno que pueden aparecer después de comparar la imagen de profundidad con TRH. Una parte se filtra usando el background y otra usando la superficie.

V. Filtrado por superficie de contorno

Desgraciadamente debido al ruido de la cámara de profundidad el suelo puede mostrar crestas y valles y cuando se compara con el nivel TRH pueden aparecer pequeñas zonas

activas. Esas zonas se filtran utilizando el valor de la superficie del contorno, dicha superficie, debe estar entre dos límites: inferior y superior para ser aceptada.

VI. Filtrado por histograma de persona

Si bien el método de limitar al área del contorno es bastante eficaz, y nos permite eliminar muchos falsos positivos de forma rápida, sin embargo, no es una condición fuerte. Para validar si un contorno de un área aceptable contiene un cuerpo se realiza el histograma de esa parte y se compara con una base de datos de histogramas de cuerpos que hemos generado a partir de la adquisición experimental. La Figura 19 es un ejemplo de un histograma de la base de datos. Se han tenido en cuenta una base de datos de más 10.000 imágenes.

El método de comparación busca las distancias entre los histogramas utilizando cuatro métodos: Correlación; Chi-Square; Intersección y Bhattacharyya [73]. Si alguna de las distancias es menor que un límite prefijado se da como válido.

Una vez hemos validado que el histograma pertenece a una persona, vamos a analizar su forma, (Figura 20), vemos que hay dos zonas diferenciadas, una superior y otra inferior separadas por un mínimo relativo. La parte superior, corresponde a la cabeza y la parte inferior al cuerpo. El mínimo relativo en la parte superior coincide con la zona del cuello (NECK) (Figura 20).

A partir de lo anterior la altura de ese mínimo el cuello puede utilizarse para realizar una máscara con los puntos que pertenecen a la cabeza,

$$Headdepth(x,y) = \\ = \begin{cases} Persondepth.Inicial(x,y) & Persondepth.Inicial(x,y) > NECK \\ Suelo & en\ \ otros\ \ casos \end{cases} \quad (29)$$

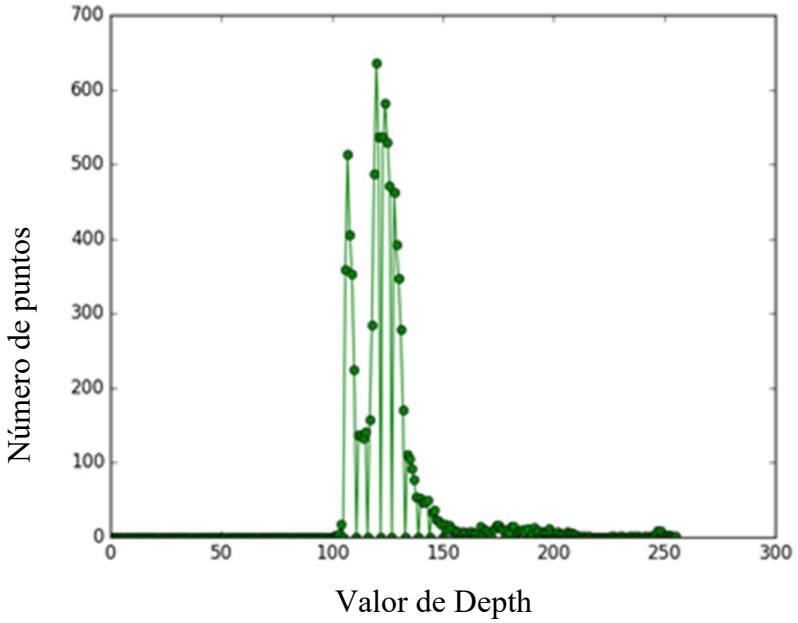


Figura 19. Histograma de otra persona en una imagen como la de la Figura 17.

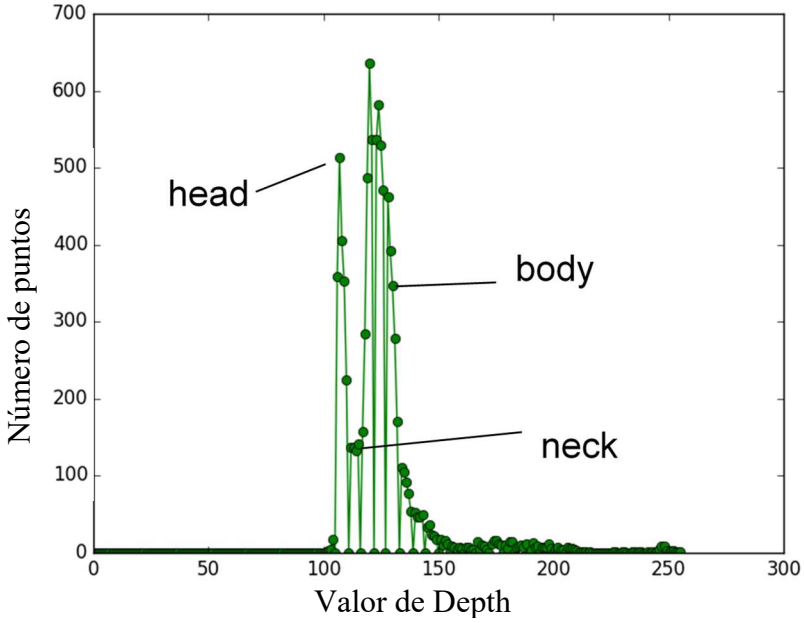


Figura 20. Partes principales (cabeza, cuello y cuerpo) del histograma de una persona.

VII.Cálculo del centroide de la cabeza

Como elemento de validación adicional comparamos el área del contorno de la cabeza con dos limites (Figura 21) (superior e inferior), que corresponden a valores determinados experimentalmente usando la misma base de datos que para determinar el valor del NECK en el apartado VI.

A continuación, se obtiene el centroide a través del cálculo del punto medio de las coordenadas de la cabeza (Figura 22),

$$Centroide(x, y, z) = \frac{1}{n}\sum_{i=1}^{n} c_i \tag{30}$$

Donde c_i son los puntos de la cabeza que hay n. Las coordenadas del centroide son las coordenadas x, y, z de la cabeza.

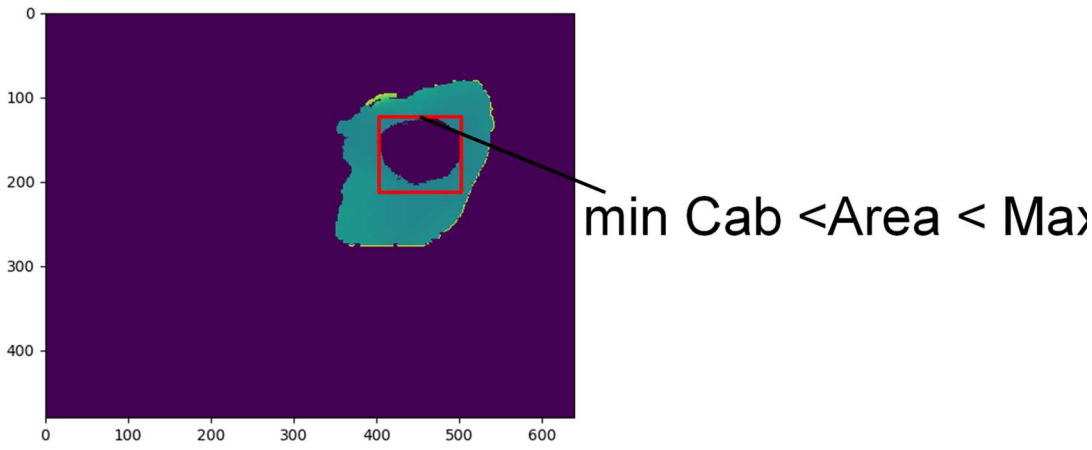


Figura 21.Cabeza y límite de superficie del contorno rectangular de la misma

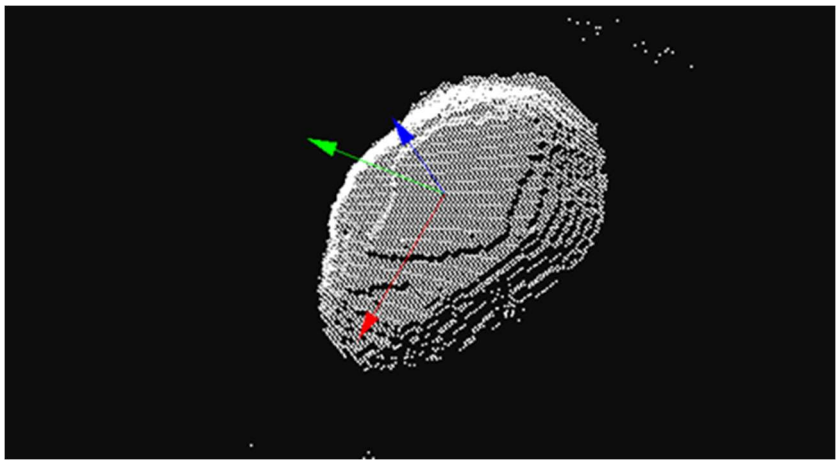


Figura 22. Nube de puntos de la cabeza y posición del centroide

VIII. Compensación de coordenadas por altura

Si bien el centro indicado corresponde a la posición central de la cabeza vista desde la cámara. Esta no es la posición real de la misma en la sala. Las coordenadas anteriores están basadas en la localización de la cabeza extraída de la imagen de profundidad, sin embargo, la posición (x, y) de los objetos obtenidos de la cámara cenital tiene un error que depende de la altura del objeto. A modo de ilustración se pone la Figura 23 que es la imagen obtenida de la cámara y como se puede ver la mesa está retrasada con respecto a sus patas, cosa que no debería de ser. En la Figura 24 se muestra la imagen compensada, donde ya no se ven las patas de las mesas porque están ocultas por la base de madera de la mesa. La compensación se obtiene mediante la aplicación de las siguientes expresiones:

$$coef = \left|\frac{z - suelo}{suelo}\right| \tag{31}$$

$$\begin{aligned} zc &= z, \\ yc &= y - \big((y - y_{cam}) \cdot coef\big) \\ xc &= x - \big((x - x_{cam}) \cdot co \end{aligned} \tag{32}$$

donde x_{cam}, y_{cam} son las coordenadas de la cámara. Para esta compensación z vale cero siendo el techo donde está la cámara y la posición (x_{cam}, y_{cam}) es en nuestro caso (640/2,480/2).

IX. Filtrado de Kalman

Adicionalmente a esta compensación, se producen errores debidos a las oclusiones de unos puntos de la cabeza, lo que produce que el centroide de la cabeza se desplace ligeramente a la zona de visibilidad de la cámara. Para compensar éste y otros problemas de ruido asociado a la medida de las cámaras realizamos un tracking de Kalman 2D de la posición [74][75].

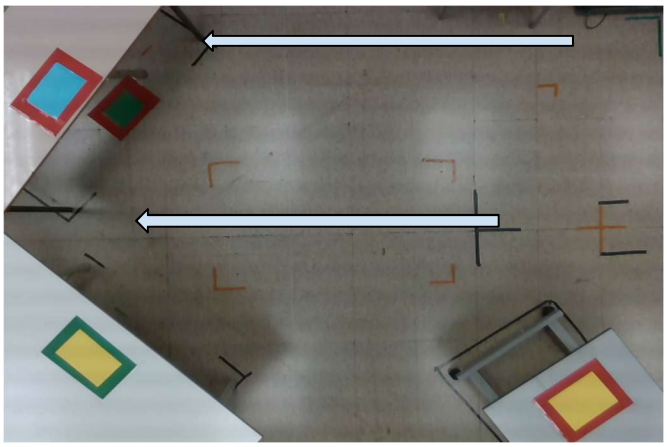


Figura 23. Imagen cenital de las mesas

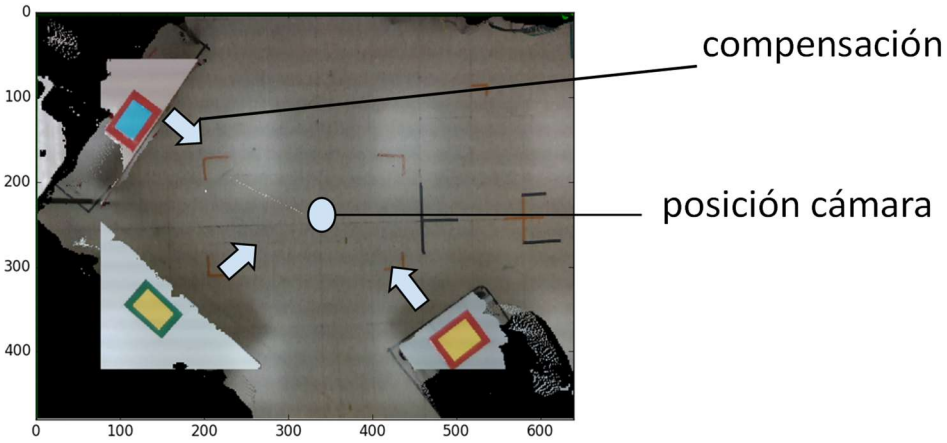


Figura 24. Imagen cenital compensada

Se utiliza un filtro de Kalman sin aceleración que solo considera variaciones en la posición, donde solo se mide la posición (x, y). Por otra parte, se considera la variación del tiempo, que es el salto de tiempo entre frames, en nuestro caso 6 frames /sec. El algoritmo

de predicción es un estimador de estados basado en el filtro de Kalman, que dadas la posición inicial y la velocidad determina la posición final y la trayectoria del objeto analizado. Las covarianzas utilizadas han sido las siguientes: E_x= $1e^{-2}$; E_y=$1e^{-2}$; E_{vx}= 5.0; E_{vy}=5.0; E_w =$1e^{-2}$ y E_h=$1e^{-2}$.

$$(fx_c\,, fy_c\,) \;=\; \text{Kalman}\,(\,x_c\,, y_c), \qquad \text{donde} \quad fz_c = z_c \tag{33}$$

4.2.2 DETERMINACIÓN DE LOS ÁNGULOS

Los ángulos para determinar son el Yaw, el Pitch y el Roll que se muestran en la Figura 25.

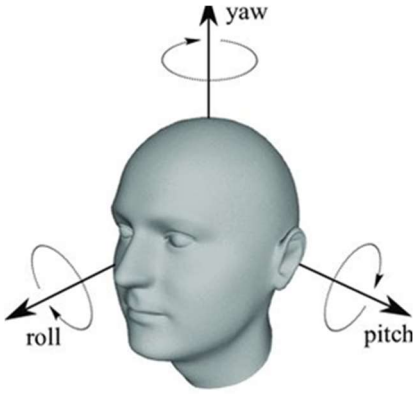


Figura 25. Ángulos a determinar[7].

4.2.2.1 DETERMINACIÓN DE ÁNGULO YAW

Para determinar el ángulo de yaw se seguirán tres pasos:

[7] Alberto Fernández Villán [online] https://www.researchgate.net/figure/Head-pose-can-be-decomposed-in-pitch-yaw-and-roll-angles_fig3_309543534. [20/07/2020]

I. Aproximación de la cabeza por una elipse

A partir del contorno de la máscara de la cabeza se busca la elipse aproximada por mínimos cuadrados como se puede ver en la Figura 26. Los ejes menor y mayor de la elipse determinan la orientación de la cabeza. El eje de la elipse marca la dirección, pero no es suficiente para determinar el sentido. En la Figura 27 se muestra la dirección en la elipse de la cabeza.

Una vez determinada la elipse además de la dirección podemos determinar la posición del centro de la elipse como se aprecia en la Figura 27.

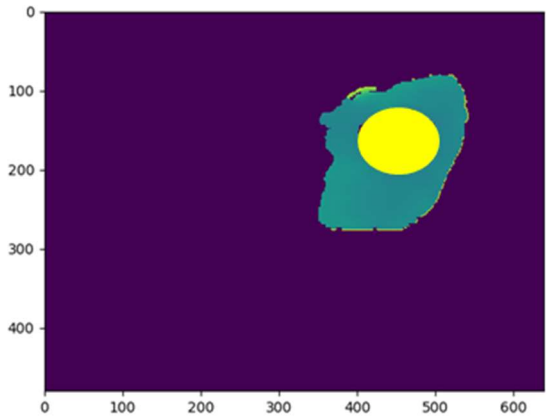


Figura 26. En amarillo se muestra la máscara de la cabeza caracterizada con una elipse

II. Método de tracking para determinar el sentido

Para determinar el sentido no se dispone de un método directo; sino que se utilizará un método de tracking, teniendo en cuenta que desde un frame al siguiente no puede cambiar 180º; es decir no puede haber un cambio de sentido de un frame al siguiente brusco. Adicionalmente debemos de conocer el sentido correcto en algunos puntos para realizar el tracking desde allí, a ese respecto utilizaremos los siguientes supuestos:

- La dirección de entrada al recinto de captación de la cámara cenital es hacia adelante (hacia la dirección de la cabeza donde hay visión).
- La velocidad usualmente dentro del recinto es hacia adelante.

- En las zonas próximas al centro de la cámara, el vector con origen en el centro del cuerpo y final en el centro de la cabeza indica el sentido.

El sentido finalmente se determina mediante un tracking llamado tracking_sentido. Este procedimiento de recuperación del sentido parte de la base de que el error en el ángulo solo puede ser ± 90º, es decir en esta fase solo tratamos de recuperar el sentido.

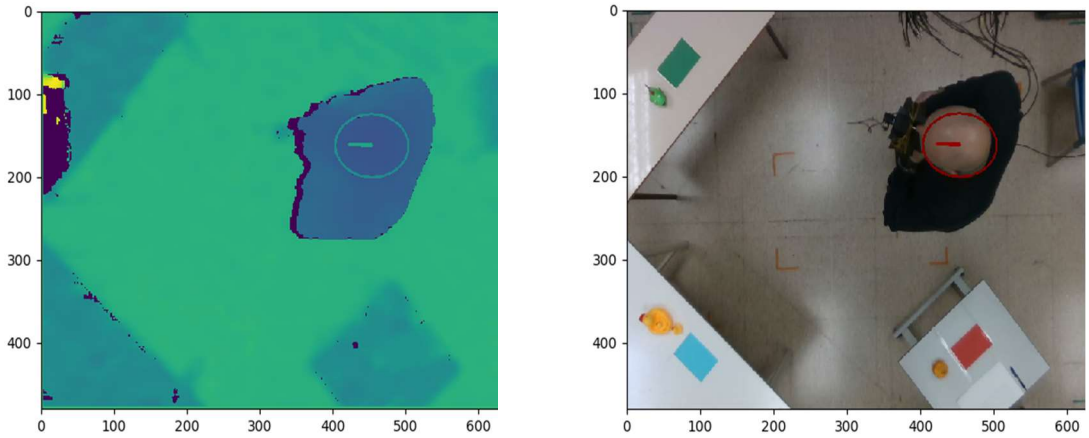
El algoritmo de tracking_sentido utilizado es el siguiente:

float ang_ori [i] con i 0..fin array ángulos iniciales

float ang_tr [i] con i 0..fin array valor final

Para realizar esta recuperación se realiza la siguiente secuencia:

ang_tr ← ang_ori



a) Imagen Depth b) Imagen RGB

Figura 27. Elipse de la cabeza con su posición del centro y su dirección

si ang_tr[0] > 90º respecto a ángulo entrada

ang_tr[0] +180

Bucle posición 1-- fin

Si ang_tr[i] > 90º respecto a ang_tr[i-1]

ang_tr[i] +180

Este algoritmo lo representamos agrupado en la siguiente ecuación:

yaw_tracking = tracking_sentido (yaw _elipse)

En la Figura 28 se puede ver en color azul los ángulos *Yaw* medidos y en naranja los ángulos que se obtendría con el método de validación que explicaremos en el próximo capítulo. Los valores en azul que se señalan con las líneas negras presentan dirección errónea (Figura 28) si se comparan con los que se obtienen con el método de validación que es el referente.

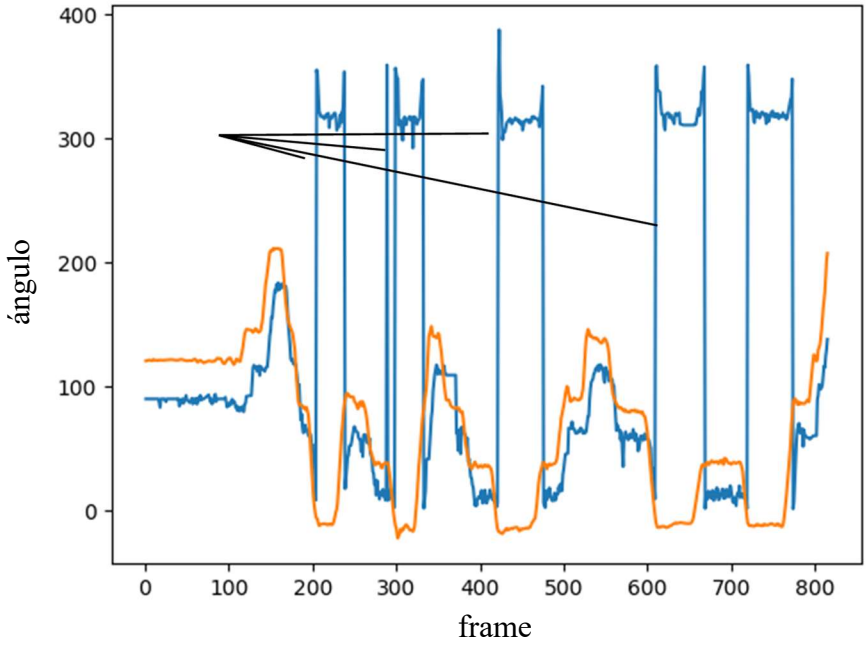


Figura 28. Ángulo Yaw con errores de sentido.

III. Filtrado de velocidad angular

Si bien el tracking nos permite encontrar el sentido del ángulo yaw, aún existe algún ruido asociado a la captura de la imagen, que mejoraremos imponiendo la condición de límite de velocidad y suavidad en el cambio de ángulo,

$$filter_{yaw} = eliptic_{dual}(yaw - tracking) \quad (34)$$

El método encontrado más efectivo ha sido un filtro pasa bajo, concretamente hemos escogido un filtro digital elíptico de fase-cero (recorrido en dirección temporal y anti-temporal).

Las características del filtro son: de orden 4, ripple 0.01, banda pasante con atenuación de 120dB, Wn=0.125. La implementación del filtro se ha realizado con la librería *scypy signal* de Python utilizando las expresiones(35) y (36). En la expresión (35) obtenemos b,a que son numerador (b) y denominador (a) del filtro IIR que se aplica en la expresión (36), donde se calcula un resultado de filtro en sentido temporal y otro en sentido antitemporal, obteniendo así una salida al filtro sin retardo, fase 0 y orden doble respecto al original.

$$\mathrm{b, a} \;=\; \mathrm{signal.\,ellip}(4, 0.01, 120, 0.125) \quad (35)$$

$$fgangulo = signal.filtfilt\ (b, a, valor, method = \text{"gust"}) \quad (36)$$

Aplicando este método se obtiene el resultado que se muestra en la Figura 29 donde los ángulos erróneos que aparecían en la Figura 28 se han corregido y sus variaciones bruscas se han suavizado.

En la Figura 30 se muestra el mismo tipo de resultado, pero para otra trayectoria diferente a la mostrada en la Figura 29. Se representa en color azul el ángulo que se calcula para cada frame filtrado según las ecuaciones (34)-(36). En color verde se muestra el ángulo real original. En el capítulo de Validación explicaremos cómo se han realizado los ensayos para disponer de valores reales de ángulos a partir de sensores IMU. Podemos ver que el error medio entre ambos resultados es de 15º que en relación a 360º es un 4%.

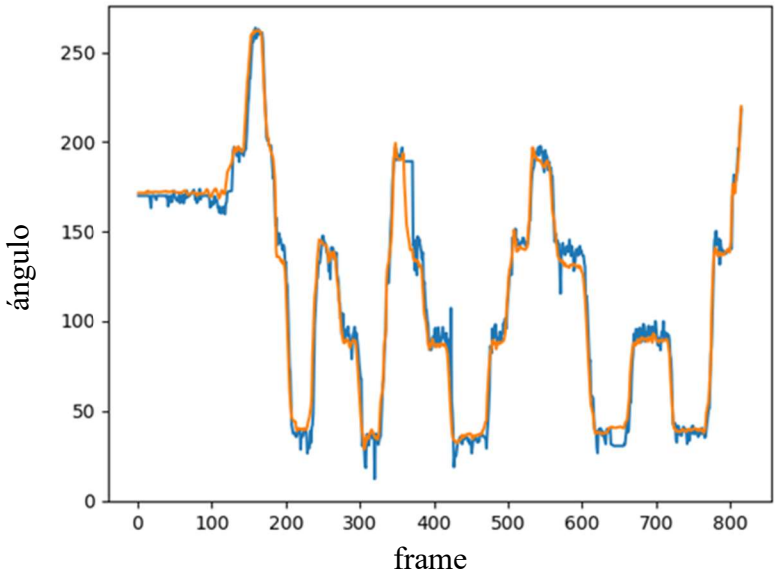


Figura 29. Ángulo Yaw con el sentido corregido

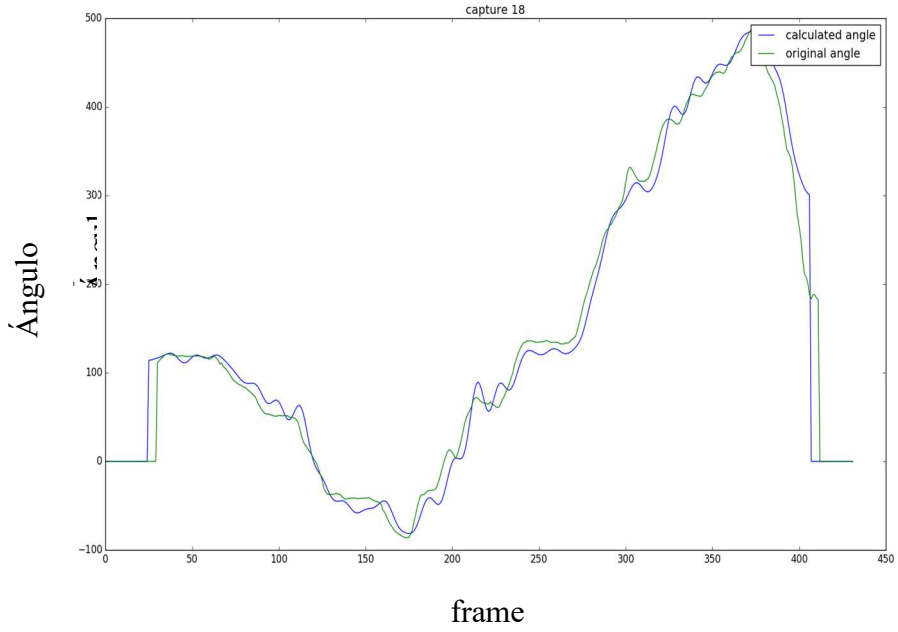


Figura 30. Ejemplo de cálculo del ángulo Yaw con filtrado de ruido incluido. En azul se representa el ángulo medido con el IMU y en verde el ángulo calculado con nuestro método.

4.2.2.2 DETERMINACIÓN DE ÁNGULO PITCH

El ángulo pitch se calcula en dos pasos:

I. Medida de la deformación del conjunto cuerpo-cabeza respecto a la vertical

El ángulo pitch mide la inclinación de la cabeza. Se han probado diferentes métodos, siendo el que presentamos el que ha proporcionado los mejores resultados. El método requiere determinar cuan deformado está el conjunto cabeza-cuerpo respecto a la vertical. Para ello es necesario determinar el centro de la parte superior de la cabeza y la altura y posición del cuerpo del fin de la deformación como se puede ver en la Figura 31**.** Con ambos datos determinamos el ángulo pitch, φ, como el ángulo entre el punto de deformación final del cuerpo y el punto superior de la cabeza con respecto al eje vertical z,

$$\varphi = \cos^{-1}\left(\frac{\overrightarrow{Z-A}}{|\overrightarrow{Z-A}|}\cdot\frac{\overrightarrow{B-A}}{|\overrightarrow{B-A}|}\right) \tag{37}$$

Mirar las magnitudes en la Figura 31. Ahora hay que determinar el punto superior de la cabeza y el punto de deformación final del cuerpo.

Para determinar el punto superior de la cabeza, determinamos qué puntos forman la parte superior de la misma y sobre ellos buscamos el centroide. El método de separación se basa en delimitar las zonas de la nube de puntos de la cabeza encontrando los bordes de ésta, utilizando el método de diferencia de normales [76]. De esta forma se separa la parte superior de la cabeza (SupC) del resto de la cabeza y se calcula su centroide,

$$\text{CentroidCS} = \text{Centroide (área de SupC)} \tag{38}$$

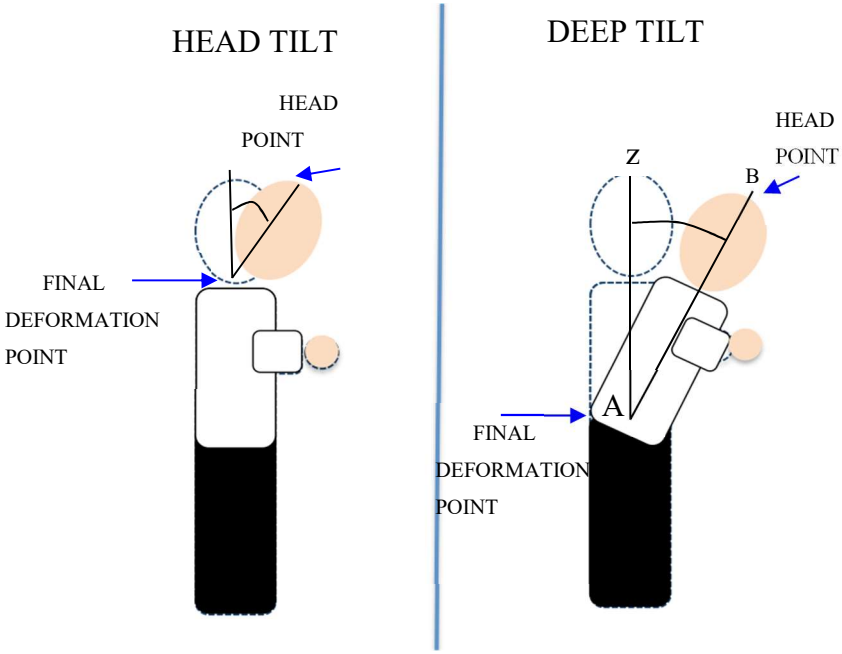


Figura 31. Imagen de ángulo pitch en relación con la deformación del cuerpo.

A continuación, se busca el punto de deformación final del cuerpo. Para ello se divide el cuerpo en 6 zonas equidistantes (de Z1 a Z6), tomando la altura de la cabeza y la altura del cuello (TRH). Se realiza la separación del cuerpo en 6 zonas y se realizan los cálculos de los centroides acumulativos:

$$CentroideZi \; = \; Centroide\;(puntos\;acumulados\;hasta\;Zi) \tag{39}$$

Se busca la evolución distancia (x,y) (componente en dirección del sentido de visionado) y se localiza las alturas Z de forma que la distancia en el plano x, y (en la dirección del sentido de visionado) sea menor que un valor dado (lim).

$$\text{Zend} = \text{Zi} \quad \text{tal que} \quad \text{dxy}(\text{Zi} - \text{Zi} - 1) \; < \; \text{lim} \tag{40}$$

A continuación, calculamos la posición del centroide del punto final. En la Figura 32 se ve una imagen captada por la cámara cenital y otra con la separación de las seis zonas en cada persona donde se usa un color para cada zona. En la Figura 33 se muestra las diferencias entre las diferentes zonas, primero es grande y luego se puede apreciar cómo llega un

momento en donde tienen más o menos el mismo valor, por tanto, no hay deformación. Esto es lo que permite determinar el ángulo pitch.

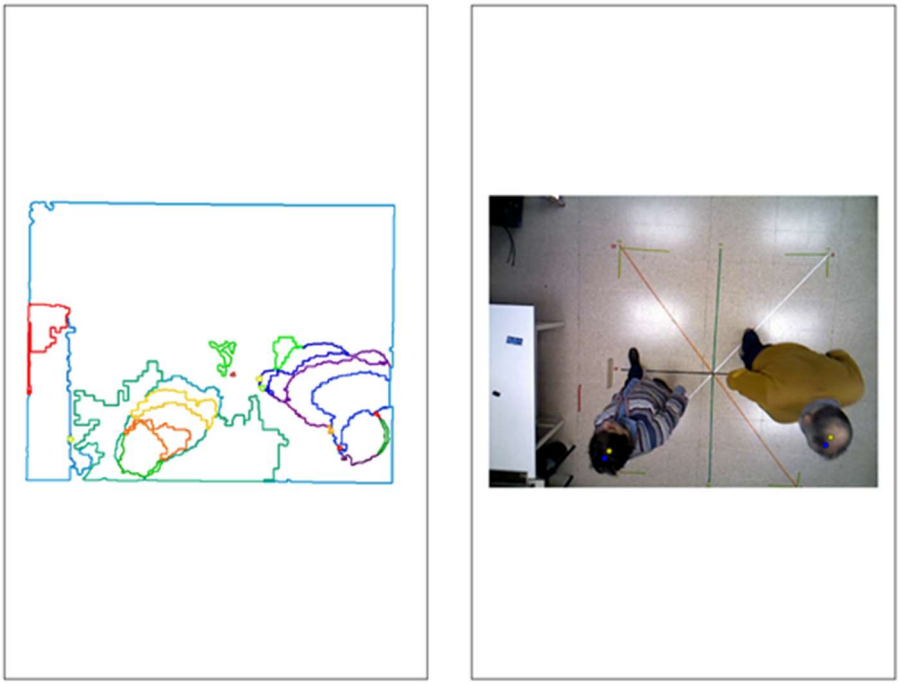


Figura 32. Imagen de la separación de las zonas, Z1 hasta Z6 del cuerpo de cada persona.

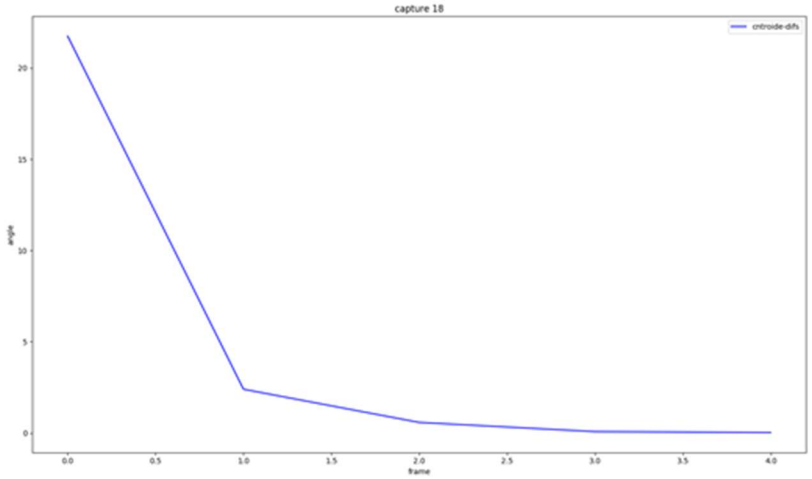


Figura 33. Distancia diferencial de la zona Z1 a la Z6 del cuerpo de una persona.

Para determinar el ángulo pitch se tiene que calcular:

$$pitch = ángulo\,(vector\,(CentroideZend -> CentroideCS)\,;\ z) \quad (41)$$

II. Filtrado de velocidad angular

Finalmente, la secuencia Pitch se filtra de forma similar al filtrado del ángulo Yaw (ecuación (34)), se eliminan los cambios de ángulo que superen la velocidad máxima de giro de cabeza, mediante un filtro de paso bajo elíptico, actuando en sentido temporal y anti-temporal para eliminar el retardo de grupo

$$\text{fPitch} = \text{eliptic\_dual (Pitch)} \quad (42)$$

El resultado del proceso de cálculo se puede ver en la Figura 34, donde se compara el valor medido mediante (IMU) en color verde con el valor calculado en azul con un error medio de 3,5º que relativo a un ángulo de 90º como límite de excursión del valor es 3,8%.

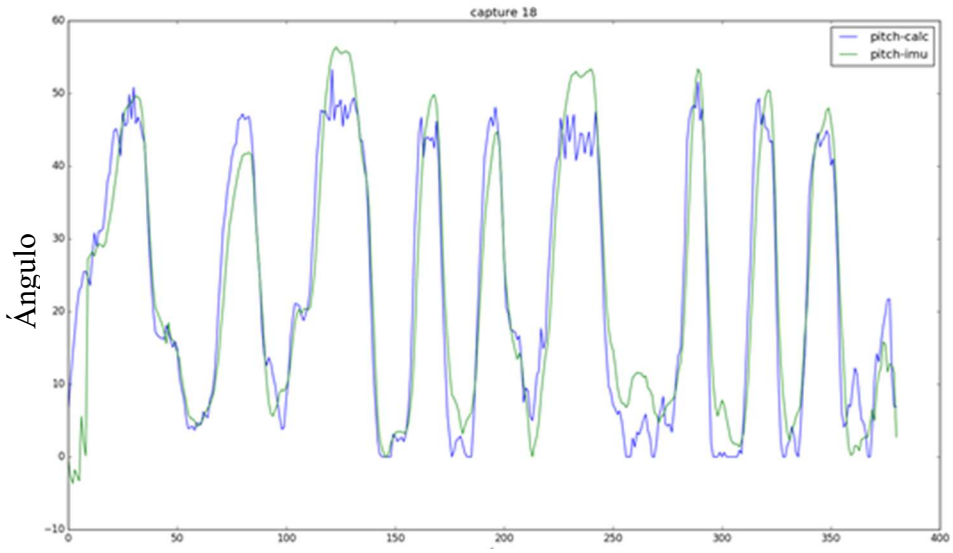


Figura 34. Comparación entre el Pitch calculado (color azul) y el medido con el IMU (color verde).

4.2.2.3 DETERMINACIÓN DEL ÁNGULO ROLL

Basado en los resultados medidos, el ángulo roll se considera cero. Los ángulos medidos de roll tienen como media 3,3º por lo que el error al considerarlo 0 es de 3,3º y sobre 90º de rango equivale a un 3,6% , por lo que considerando 0 el ángulo de roll estaremos dentro de los errores de estimación de los otros ángulos.

4.2.3 CÁLCULO DFOA Y VFOA

Los pasos que seguir para el cálculo de la DFOA son:

1. Cálculo de la trayectoria: ya realizada al calcular la posición de la cabeza y sus ángulos en los apartados anteriores
2. Determinación de la zona de análisis: para reducir tiempo de cálculo se determina una zona de análisis
3. Cálculo de la DFOA

4.2.3.1 ZONA DE ANÁLISIS

La materialización de la zona de análisis debe de tener la posibilidad de describir mediante una técnica probada un entorno donde se mueven las personas y donde se puedan describir todos los objetos susceptibles de tener algún interés.

La zona de análisis consiste en una nube de puntos que contiene, para cada uno, la posición (x, y, z) del punto, y el vector normal a la superficie en la que se haya. La nube contiene solo los puntos de paredes, suelos, superficies y objetos de interés. No contiene puntos pertenecientes al aire para aligerar el cálculo.

Para ello utilizaremos la librería PCL (point cloud library) [77]. Concretamente la nube de puntos con sus normales, *Point Normal,* que nos define el entorno y la localización de los objetivos. A esta nube de puntos la llamamos **AZ (*analysis zone*).**

$$\text{AZ(analisis zone)} = \text{PCL cloud (Point Normal)} \tag{43}$$

Generamos la zona de análisis en dos zonas diferenciadas:

- **Superficies y objetos captados por la cámara:** Contiene las superficies y objetos capturados por la cámara cenital. Es la zona donde se captará también a las personas con las que se calculará la trayectoria corregida en posición y la VFOA. La densidad de puntos (número de puntos por metro cuadrado) de esta nube es alta y determinada por la resolución de la cámara utilizada.
- **Paredes y suelos:** Las paredes de la sala suelen estar fuera del alcance de la cámara cenital, así como gran parte del suelo de la sala, pero es fundamental incluirlos en la nube de puntos por si hay superficies de interés en ellas La densidad de puntos en estas zonas es baja, ya que no se genera con la cámara cenital, sino con un escáner extra, o bien se construye manualmente, como ha sido el caso de nuestros ensayos de validación.

Por otra parte, el algoritmo de procesado tiene un orden de procesado O(nlogn) con lo que se puede optimizar el cálculo reduciendo el número de puntos. La imagen de Figura 35 es un ejemplo de la zona de análisis.

Se puede observar que la densidad de puntos es mucho menor en el suelo y en las paredes. Hay que hacer notar que, para poder distinguir bien el dibujo en dos dimensiones, se ha ordenado los puntos de la pared en filas horizontales. Si se hubieran representado todos los puntos, como se ha hecho en el suelo, sería imposible distinguir las paredes y el suelo. En la nube de puntos de las 3 paredes se pueden ver un cartel de color en cada una de ellas. En total son 3 carteles de color: verde, rojo y azul. Estos carteles serán objetivos de la VFOA, aunque estén fuera de la zona captada por la cámara cenital. Y el método de cálculo debe dar valores de VFOA en esos puntos.

La zona captada por la cámara se observa de color negro debido a que la densidad de puntos es muy elevada. En esta zona se ha incluido los puntos de las mesas con sus carteles (amarillo, magenta y azul) captados por la cámara, y la proyección de los mismos al suelo (Figura 35).

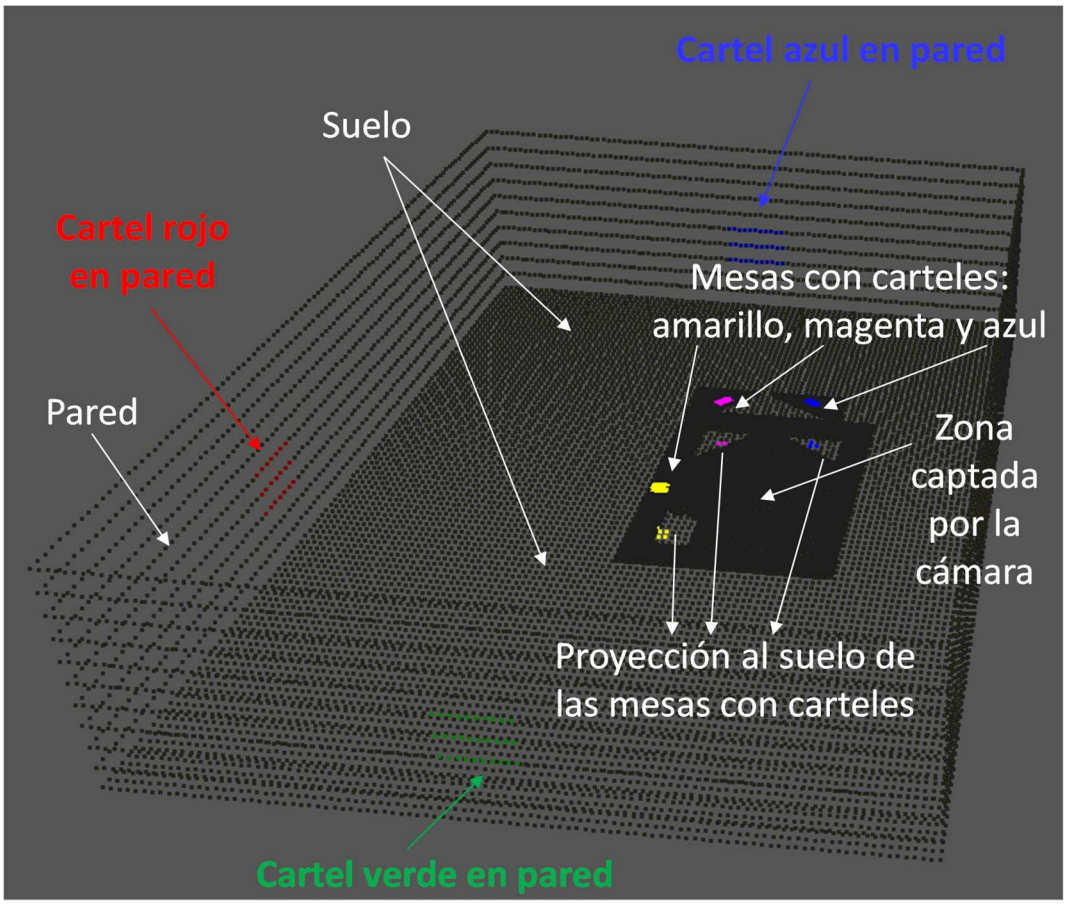


Figura 35. Zona de análisis, se puede ver la zona que capta la cámara, paredes, suelo y targets.

4.2.3.2 CÁLCULO DE DFOA

De las expresiones indicadas en la sección del método DFOA para cada elemento de la trayectoria se calculará con la siguiente expresión

$$DFOA(P) = K \cdot F_{ángulo} \cdot \frac{1}{d^2} \cdot cos\theta_N \tag{44}$$

Para el cálculo de la *DFOA* se utilizará una nueva nube de puntos de la misma dimensión que *AZ*. Tendrá los mismos puntos, y en cada uno se guardará la posición (x,y,z) y una magnitud escalar que representará la *DFOA* en ese punto.

En este caso utilizaremos la nube de puntos *Point Intensity* de la librería PCL (*point cloud library*) [77]. Concretamente la nube de puntos con sus normales *Point Normal* que nos define el entorno y la localización de los objetivos. A esta nube de puntos la llamamos **IZ (*intensity zone*).**

$$\text{IZ(intensity zone)} = \text{PCL cloud(Point Intensisy)} \tag{45}$$

La razón de no utilizar una única nube de puntos que contenga posición (x,y,z), vector normal, y valor de DFOA, es que en las funciones de las librerías PCL utilizan nubes de las características reseñadas.

Para cada punto de la trayectoria del individuo se calcula la DFOA en todos los puntos de la nube de la zona de análisis (*AZ)*. Para ello:

1. Se toma el vector normal del punto de la nube
2. Se calculan los ángulos de la cabeza del individuo (como se ha indicado anteriormente)
3. Se calcula la DFOA según la expresión (44).

En este proceso es necesario transformar las coordenadas de la sala, en las que están los puntos de *AZ*, en coordenadas del ojo del individuo. Esto es así porque hay que calcular ángulos de visión y distancia del ojo (o cabeza) al punto de la nube de la zona de análisis para poder determinar el interés de la persona. La matriz de transformación utiliza los ángulos yaw, pitch y roll y la traslación desde el origen de la sala a la posición de la cabeza. Pueden encontrarse fundamentos de estas matrices en la siguiente referencia [78]:

$$\text{Mt} = [\text{R}][\text{T}] \tag{46}$$

Esta matriz Mt se aplica para transformar la DFOA de la sala a la referencia del ojo. Transformamos las nubes de puntos, utilizando directamente las API del PCL (PCL:transformPointCloudWithNormals). Posteriormente se realiza la transformación inversa, M_t^{-1} para volver a las coordenadas de la sala y poder sumar con los valores anteriores.

Los resultados parciales de la *DFOA* los vamos almacenando en la nube de puntos *de la zona de intensidad (IZ)*. En ella, la posición de los puntos (x,y,z) es la misma que en *AZ*. Lo que varía es que, en lugar de almacenar el vector normal, almacenamos el valor parcial de *DFOA*. Al final del proceso, en la nube de puntos IZ tendremos la *DFOA* en cada una de sus posiciones.

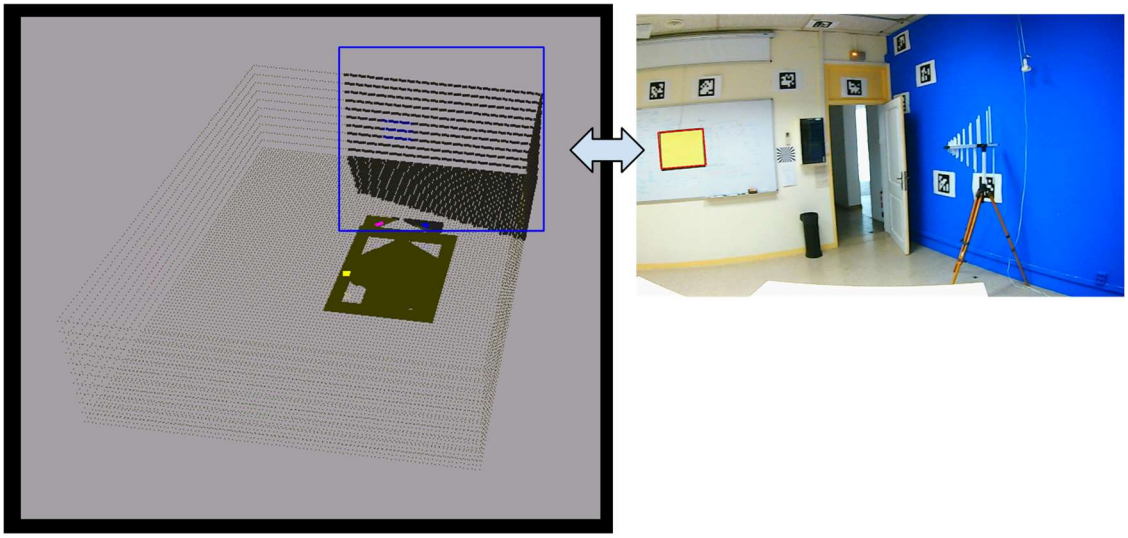
Hay que notar que para nuestro caso no requerimos manejar oclusiones, puesto que solo hay un sujeto en el test y los objetos a analizar están en zonas no ocultas por ningún otro objeto. En el caso más genérico se deberían de considerar, si bien esto no introduce demasiada complejidad en el proceso, solo tiempo de cálculo y memoria.

El cálculo final de la intensidad DFOA se puede observar en la Figura 36 en donde se muestra la visión real de un individuo (Figura 36 b)) y la DFOA en ese mismo instante (Figura 36 a)).

En la representación gráfica de la DFOA (Figura 36 a)), la imagen que se ve a la derecha se corresponde con la mayor densidad de foco de atención dado que en ese momento es hacia donde se está mirando. Eso se corrobora exactamente con la imagen de lo que está viendo o enfocando en el individuo (Figura 36 b).

Hay que hacer notar que el póster real es de color amarillo, mientras que en la representación gráfica de la *DFOA* se ha optado por dibujarlo en azul. Pero nos estamos refiriendo al mismo póster. Si calculáramos el área en pixeles del póster en azul de la gráfica de la *DFOA* (Figura 36 a)), se correspondería con el área del poster amarillo de la imagen real del individuo (Figura 4-1 b)).

Nótese que en el cálculo *DFOA* se ha utilizado la función del ángulo (7), relacionado con el comportamiento de la cámara frontal utilizada para validar los resultados. En el capítulo 5 de validación se explicará más extensamente cómo se han validado los resultados.



a) Representación gráfica de la DFOA b) Visión real del ojo del individuo

Figura 36.Comparación con la imagen capturada por la cámara en esta posición de la trayectoria.

4.2.3.3 CALCULO DE VFOA

El cálculo de la VFOA se realizará sumando las diferentes DFOA,

$$VFOA(P) = N \sum_{C} \sum_{T_r} DFOA(P) \tag{47}$$

La dificultad de este proceso es la búsqueda de los mismos puntos para realizar esta suma; pensemos que el conjunto de puntos no tiene estructura y por tanto de no realizar ninguna acción tenemos un algoritmo de $O(N^2)$ de muy lenta ejecución. Para acelerar el proceso se genera dentro de la nube de puntos de Intensidad (IZ), una estructura octree [8] que convierte el algoritmo de $O(N^2)$ a O(NlogN) para que el algoritmo resulte manejable.

[8] http://pointclouds.org/documentation/tutorials/octree.html

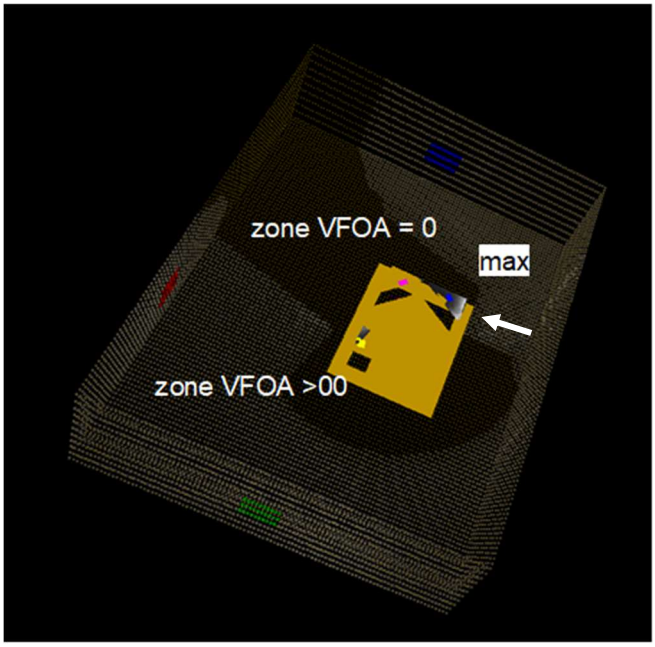


Figura 37. VFOA de una trayectoria

En la Figura 37 se representa la magnitud de la atención en blanco en cada punto siendo la más intensa blanca máximo (ver flecha blanca que apunta a una zona blanca) y la negra atención nula. La zona con atención nula indica que el sujeto no ha mirado en absoluto a esa zona. La Figura 37 muestra la VFOA de una trayectoria, se puede ver que hay zonas con atención muy baja o cero y otras con el círculo con atención máxima.

4.3 IMPLEMENTACION 2D

Para la implementación en 2D no se puede utilizar la técnica usada en 3D. El método 2D no tiene tanta información y en general para reconocer la cabeza entrenan con una base de datos con imágenes que contienen cabezas y otras no. De ese modo luego pueden determinar con otras imágenes que probabilidad tienen de tener una cabeza.

Para determinar qué técnica era más adecuada para nuestro objetivo, detectar la cabeza con imágenes cenitales, se han probado dos técnicas que nos han interesado revisando el

estado del arte. Seguidamente se explica en qué se basan, se proporciona los resultados obtenidos y se llega a una conclusión[9].

4.3.1 DETECCIÓN DE LA CABEZA.

4.3.1.1 MÉTODO SVM Y HOG

Para la detección de la cabeza se ha tomado como primera opción un clasificador de máquinas de vectores de soporte, SVM ("*Support Vector Machine*") [79] utilizando como características de las imágenes de los histogramas de gradientes orientados, HOG ("*Histogram of Oriented Gradients*") descritas en [80][81] para la clasificación de los objetos. Para este método se ha diseñado un clasificador SVM como el descrito en [44] con el objetivo de obtener resultados similares a los obtenidos por el autor, esto es a partir de los descriptores HOG con el Clasificador SVM clasifica si ese pixel pertenece a la cabeza o no.

Para una función bidimensional f(x,y) el gradiente se expresa como:

$$\nabla f(x,y) = \left[\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y}\right] \tag{48}$$

El gradiente es un vector perpendicular a los contornos fuertemente marcados en la imagen. Este gradiente se calcula mediante filtros lineales para la dirección en el eje X y en el eje Y por separado realizando una convolución discreta en cada pixel con máscaras como las que se indican en las ecuaciones (49) y (50). La ecuación (49) se utiliza para el cálculo de la derivada horizontal y la (50) para el cálculo de la derivada vertical.

$$G_x = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} \tag{49}$$

[9] Los apartados pertenecen al TFG titulado "Generación de algoritmos de reconocimiento de personas captadas cenitalmente" del autor Eduardo Bernal Pérez mientras trabajaba en la empresa Venco al igual que el autor de esta tesis que lo supervisó.

$$G_y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \quad (50)$$

Una vez obtenidos los valores G_x y G_y se puede calcular el módulo y la fase del gradiente. Para calcular el módulo del gradiente:

$$|\nabla f(m,n)| = \sqrt{G_x^2 + G_y^2} \quad (51)$$

El módulo indica cuánto varía una función en un punto, en nuestro caso en el punto [m, n] de la imagen. Para calcular la fase del gradiente:

$$\emptyset(m,n) = arctg\left(\frac{G_y}{G_x}\right) \quad (52)$$

La fase indica la orientación y la dirección de la variación del gradiente en un punto determinado.

Para generar el Histograma de Gradientes Orientados se analiza una imagen con ventanas deslizantes, cada ventana se divide en bloques del mismo tamaño con cierto solape entre ellos. Para cada píxel de cada bloque se calcula el gradiente y para cada valor calculado se genera un peso. Con cada valor se genera un histograma aproximando según la cantidad de valores que se quiere representar en el histograma. En esta tesis se ha hecho el análisis de las imágenes con ventanas deslizantes de 64x64 pixeles y bloques de 32x32 pixeles con un solape del 50% entre sí para un total de 9 bloques en cada ventana. Cada bloque genera un histograma con 9 valores y finalmente los histogramas se combinan para generar un único vector de características para caracterizar cada ventana con un vector unidimensional. En la Figura 38 se observa el proceso de forma gráfica.

Un clasificador SVM es un clasificador binario, es decir, clasifica entre dos clases, la positiva y la negativa. Dado un conjunto de muestras de dos clases representadas en el espacio el algoritmo SVM busca una recta o un hiperplano que separa las muestras negativas de las positivas maximizando el margen de distancia entre ellas.

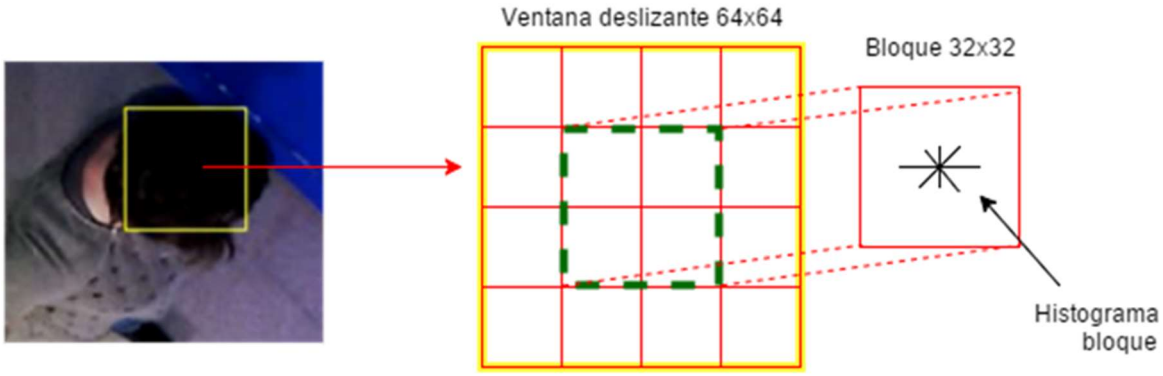


Figura 38.Como se obtienen los histogramas de gradiente orientados.

Como se observa en la Figura 39, un conjunto de muestras se puede separar con más de un hiperplano o recta, el objetivo del algoritmo SVM es decidir entre los hiperplanos o rectas posibles aquel que maximice el margen de distancia entre las dos clases.

Para entrenar el clasificador SVM es necesario un set de imágenes etiquetadas con la clase, extraer las características de cada una y a partir de los vectores de características del conjunto de imágenes de entrenamiento obtener el hiperplano que divide las clases (Figura 40).

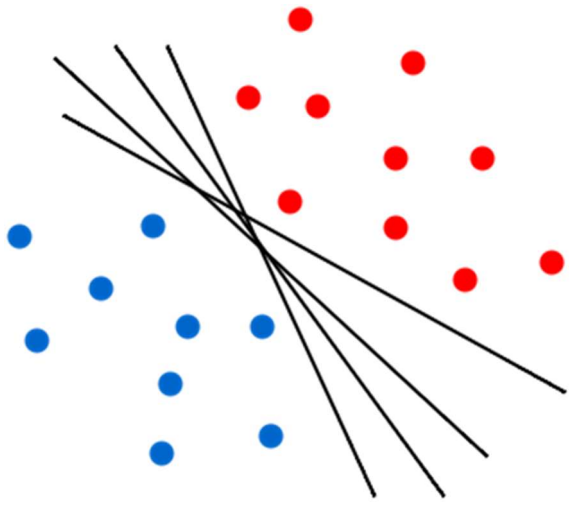


Figura 39. Esquema del cálculo de SVM.

Existen distintos tipos de clasificadores SVM, en este proyecto se ha implementado un clasificador SVM polinómico [82] que encuentra una curva y no una recta para separar las clases. Este tipo de clasificador es mejor, ya que separa mejor los datos en el caso en que las muestras se encuentren menos dispersas (Figura 41). Durante el desarrollo de este proyecto el clasificador SVM polinómico obtuvo mejor respuesta a los datos.

El clasificador implementado fue entrenado con 2206 imágenes positivas y 5441 imágenes negativas. El elevado número de imágenes negativas se debe a la cantidad de detecciones erróneas que se han obtenido en los diferentes clasificadores entrenados. El tiempo empleado en la fase de entrenamiento fue de 5 a 7 minutos.

Para la primera fase de test se han analizado 202 imágenes de 64x64 pixeles, 101 positivas y 101 negativas; estas imágenes se han obtenido utilizando un programa de edición manual realizado a este efecto. Para poder analizar los resultados obtenidos se genera una tabla de la verdad, Tabla 1, con los resultados de la clasificación.

Una vez se ha entrenado el clasificador se testeo para comprobar su correcto funcionamiento. Para este proyecto se han realizado dos fases de test, una primera con un set de imágenes positivas y negativas y la segunda fase con varias secuencias de video:

- La cantidad de verdaderos positivos es de 93 imágenes (VP).
- La cantidad de falsos positivos es de 1 imagen (FP).
- La cantidad de verdaderos negativos es de 100 imágenes (VN).
- La cantidad de falsos negativos es de 8 imágenes (FN).

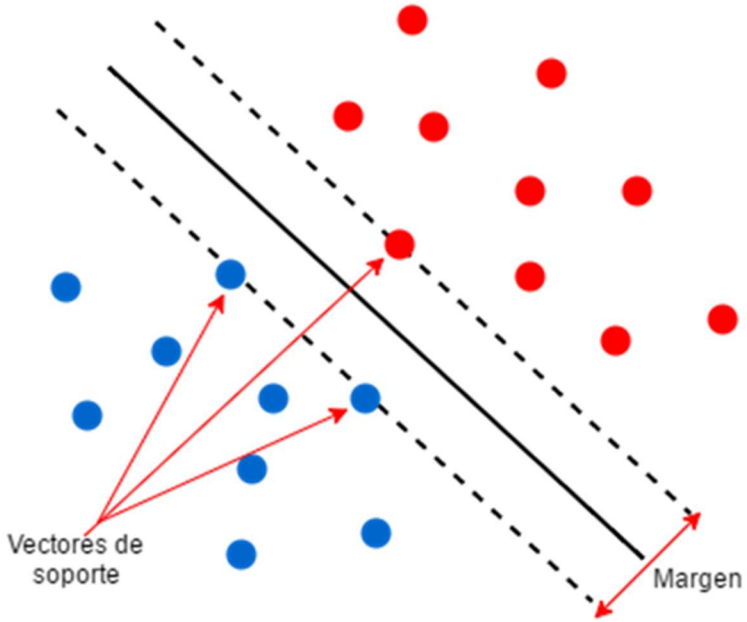


Figura 40. Clasificador lineal basado en el margen máximo a partir de los vectores de soporte.

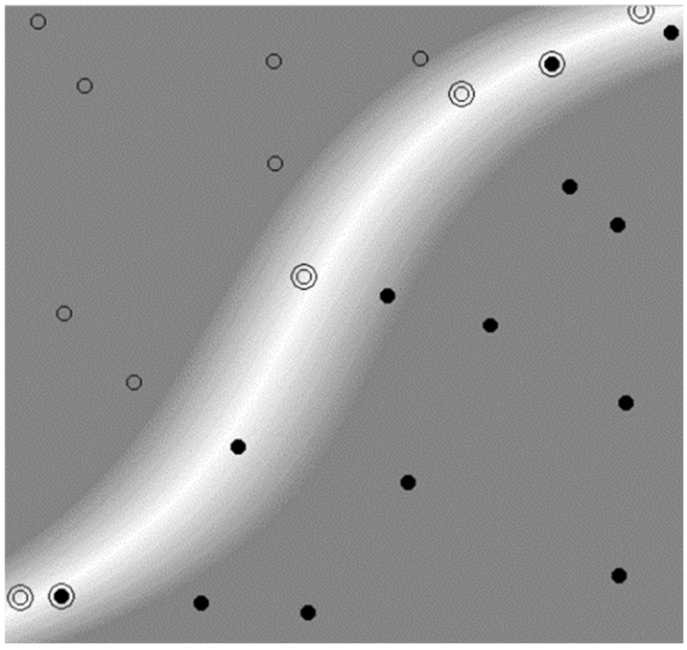
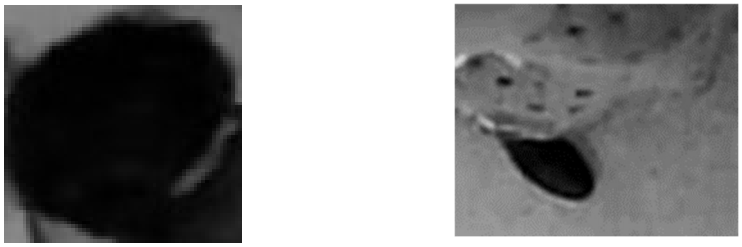


Figura 41. Clasificador SVM polinómico.



a) b)

Figura 42. Ejemplo de imágenes: a) positiva, b) negativa

Tabla 1.Resultados de la clasificación SVM polinómica.

		Predicción	
		Positivos	**Negativos**
Ground truth	**Positivos**	93	8
	Negativos	1	100

Con los valores de la Tabla 2 se procede a calcular los diferentes parámetros que sirven para analizar el comportamiento de un clasificador y son:

- Exactitud, evalúa la proporción de imágenes clasificadas correctamente respecto al total de imágenes.

$$Exactitud = \frac{VP + VN}{VP + FP + VN + FN} \tag{53}$$

- Precisión, evalúa la calidad de la respuesta del clasificador.

$$Precisión = \frac{VP}{VP + FP} \tag{54}$$

- Sensibilidad o *Recall*, evalúa la eficiencia en la clasificación de todos los objetos que son de la clase.

$$Recall = \frac{VP}{VP + FN} \tag{55}$$

- F-score, media armónica entre la precisión y el *Recall*.

$$F\text{-}score = 2 * \frac{Precisión * Recall}{Precisión + Recall} \tag{56}$$

- *Miss rate*, evalúa la cantidad de no aciertos.

$$Miss\ rate = \frac{FN}{VP + FN} \tag{57}$$

El objetivo final es poder realizar una comparación entre los clasificadores analizando los valores escogidos y el comportamiento que un observador percibe. Los resultados obtenidos para el clasificador SVM se muestran en Tabla 2Para evaluar el efecto que puede tener el solapamiento entre la ventana escogida, se ha escogido ventanas de 64x64 pixeles y se han hecho tres pruebas referentes al solapamiento entre las ventanas que recorren las imágenes. En el primer caso se realizó el análisis con ventanas que tenían un 50% de solape (Figura 43 a)), en el segundo caso con ventanas que tenían un 75% de solape (Figura 43 b)). Por último, una tercera prueba con ventanas con el máximo solape posible 100%, es decir, una ventana por pixel (Figura 43 c)). En este caso el programa era demasiado lento (más de 60 segundos por imagen) y los pocos resultados que se obtuvieron no mejoraban respecto a los otros solapes probados, por ello se desestimó este tipo de solapamiento y de aquí en adelante no se muestran los resultados.

Una vez realizadas las pruebas y analizados los resultados podemos ver que, según los resultados numéricos, la mejor forma de detectar cabezas con un clasificador SVM y características HOG es realizando una búsqueda con ventanas deslizantes de 64x64 que tengan entre sí un 75% de solape.

Tabla 2.Resultados del test clasificador SVM con las imágenes test.

	SVM
Verdaderos positivos	93
Falsos positivos	1
Verdaderos negativos	100
Falsos negativos	8
Exactitud	0,955446
Precisión	0,989362
Recall	0,920792
F-Score	0,953846
Miss Rate	0,079207921

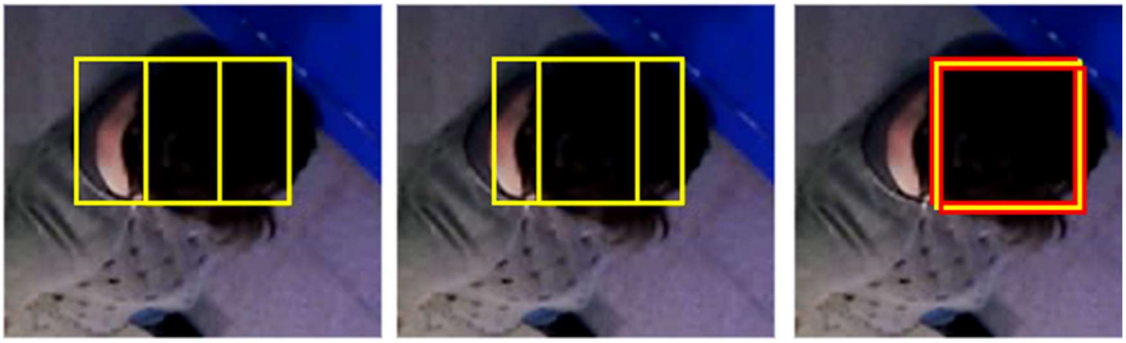


Figura 43. Pruebas con diferentes solapamientos de la ventana de 64x64: a) 50%, b) 75% y c) 100%.

Si nos fijamos en el F-score las ventanas con 75% de solape ofrecen mejores resultados numéricos. Se ha observado que cuando se establece un mayor solape entre las ventanas la cantidad de cabezas no detectadas es muy baja, esto es bueno ya que no se pierde el rastro de las cabezas, pero por otra parte la cantidad de falsos positivos es mayor y esto comporta problemas si luego se quiere analizar hacia donde se está mirando. Durante las pruebas se ha observado que la cantidad de falsos positivos cuando no aparece ninguna cabeza en la imagen es alta, por tanto, se debe encontrar una forma de mejorar estos resultados. Una posible solución sería realizar un análisis de movimiento de la imagen previa a la detección.

Teniendo en cuenta que nuestro objetivo es poder detectar las cabezas lo mejor posible y en el menor tiempo posible, finalmente hemos optado por el clasificador que analiza la imagen con ventanas con un solape del 50%, ya que se obtiene menor cantidad de falsos positivos y el tiempo de procesado es 3 veces menor.

En la Figura 44 se muestra como a medida que pasan las personas por la zona de captación se va reconociendo la cabeza. Se aprecia como a veces aparecen falsos positivos (Figura 44 d)) o no detecta la cabeza (Figura 44 c)).

4.3.1.1 MÉTODO BOOST CASCADE LBP

Para la detección de cabezas se ha tomado como segunda opción la búsqueda y detección mediante el clasificador Boost Cascade [83], una serie de clasificadores en cascada basados en Boosting [84], utilizando LBP ("*Local Binary Pattern*") descritas en [85] cómo características de los objetos a clasificar. Para este método se ha entrenado el sistema de clasificadores en cascada de OpenCV generando los ficheros necesarios para ello.

El entrenamiento y el test se realizó con las mismas muestras que se utilizaron en el apartado 4.3.1.1. donde se utilizaba el clasificador de máquinas de vectores de soporte, SVM y los histogramas de gradientes orientados (HOG).

Las características LBP ("*Local Binary Pattern*") se describen en el año 2007 en la ICB ("*International Conference on Biometrics*") [85] como un tipo de enfoque para el reconocimiento de caras. Ahora se está utilizando en otros ámbitos. LBP es un descriptor de texturas cuyo objetivo es escanear la imagen con una subventana escalable desde la cual se

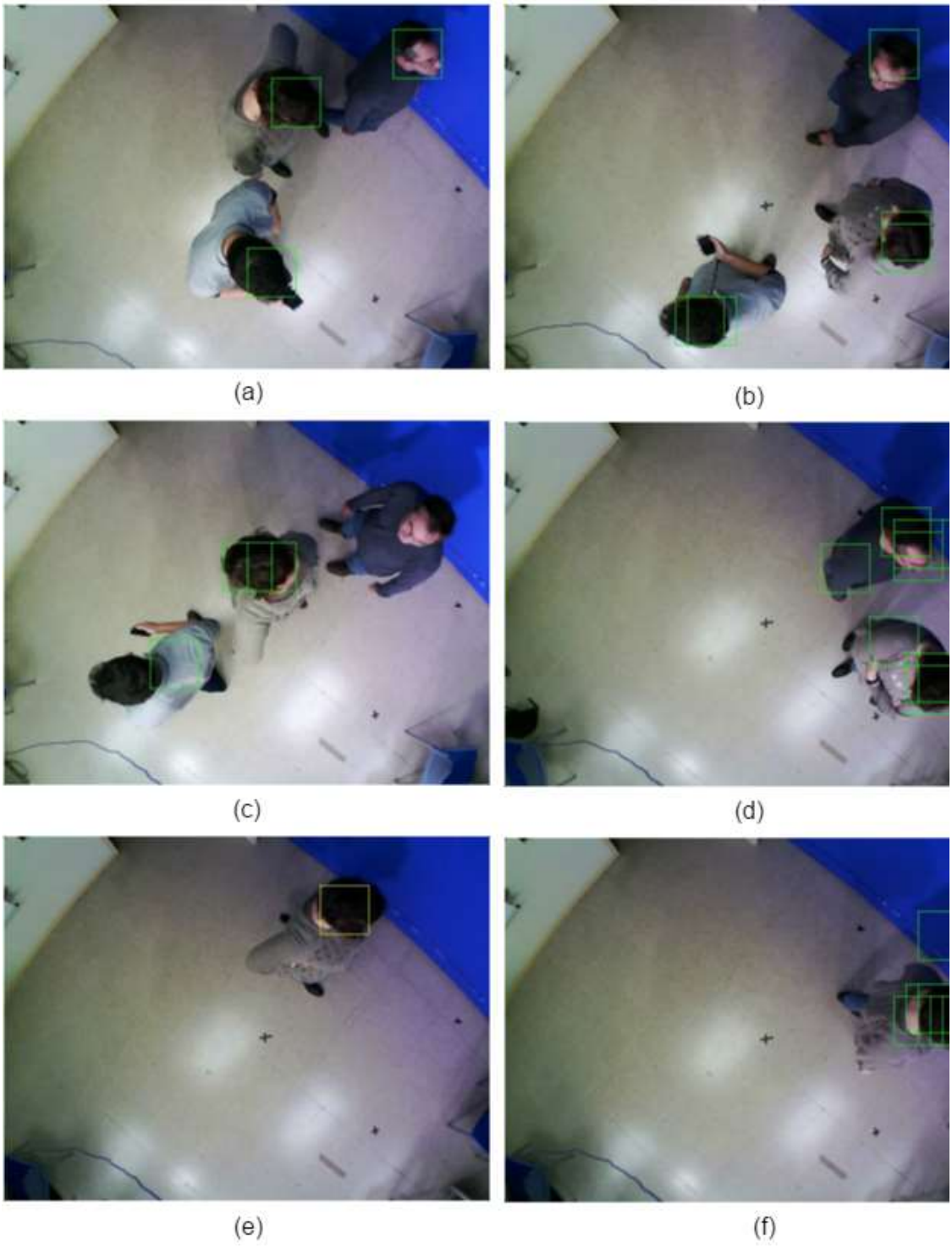


Figura 44. Imágenes donde se puede apreciar cómo se va detectando la cabeza en la trayectoria de las personas.

obtienen los histogramas de Patrón binario local (LBP) para describir las características locales de esa imagen. En su forma más simple, LBP analiza la imagen con ventanas que tengan un tamaño divisible entre 16, ya que la ventana se ha de dividir en bloques de 16x16 pixeles. Para cada píxel del bloque de 16x16 pixeles se analizan los 8 vecinos que tiene y se establece un umbral de forma que si un vecino es menor que el píxel que se está analizando se pondrá un 0 y si es mayor o igual se pondrá un 1. El análisis de los vecinos ha de ser siempre en el mismo orden para cada píxel de la imagen (Figura 45).

Una vez tenemos una sub-ventana con los valores de '1' y '0', se genera un número binario de 8 bits siguiendo el orden de análisis, este número se codifica en decimal y se genera un histograma con cada valor calculado. Este histograma será el vector de características de la imagen (Figura 45).

El clasificador Boost Cascade se genera realizando una suma de un número definido de clasificadores débiles lineales que juntos crean el clasificador fuerte. El algoritmo se describe en [83] utilizando filtros Wavelet Haar para la detección de caras, pero en esta tesis se genera el clasificador utilizando las características LBP debido a la velocidad de entrenamiento y a los resultados obtenidos. Durante el desarrollo del proyecto se realizó un entrenamiento de un clasificador con características Wavelet Haar para crear la cascada de clasificadores. Este entrenamiento tardó más de 48 horas y los resultados eran peores que utilizando las características LBP, por este motivo se desestimó.

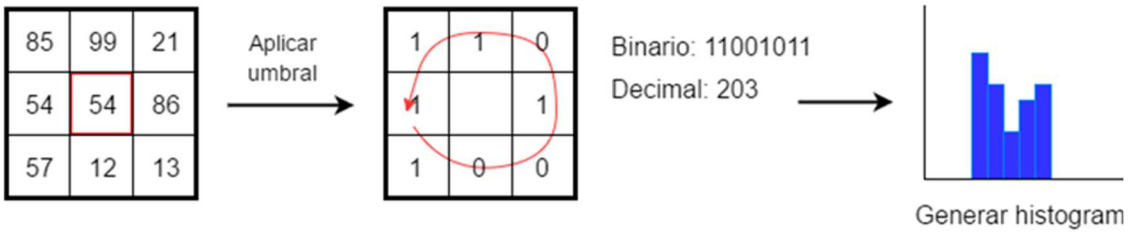


Figura 45. Ejemplo de obtención del histograma usando el método de LBP

El clasificador Boost Cascade se forma con una serie de clasificadores débiles lineales entrenados sin supervisión implementando el método de Boosting. Para entrenar el sistema es necesario tener una base de datos correctamente etiquetada. Un sistema de clasificación

en cascada se divide en etapas, cada etapa es un clasificador débil y el número de clasificadores débiles indica el número de etapas que forman el clasificador. Para considerar un clasificador débil como adecuado ha de tener una exactitud mayor que el 50% de forma que su clasificación sea mejor que una suposición aleatoria.

Con un conjunto de datos como los que se indican en la Figura 46, se puede observar que es necesario más de un clasificador lineal para poder separar los datos lo más precisos posible. Cada muestra tiene una etiqueta indicando la clase y un peso indicando la importancia de la muestra.

En la primera fase se busca un clasificador débil que separe los datos con una exactitud mayor al 50%, dejando pasar como buenas la mayor cantidad de imágenes positivas posibles.

En la Figura 47 podemos observar que una vez se estima el primer clasificador lineal se reajustan los pesos de cada muestra clasificada incorrectamente para que el siguiente clasificador obtenga mejores resultados. Para obtener el clasificador débil se realizan varias iteraciones hasta obtener un clasificador lineal débil adecuado.

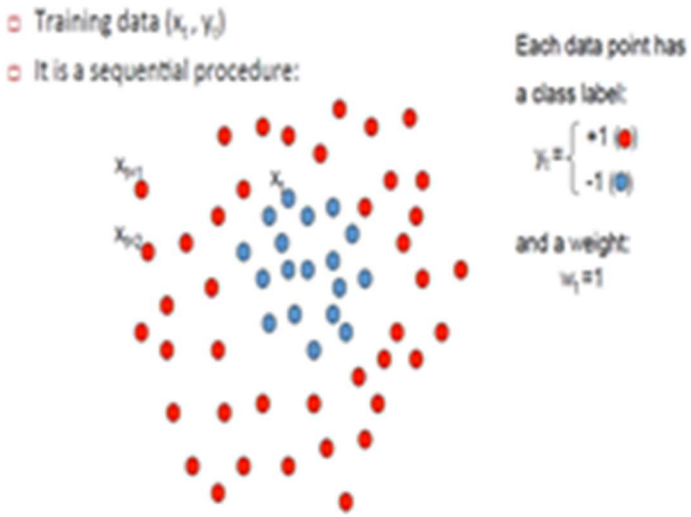


Figura 46. Datos etiquetados sin separar.

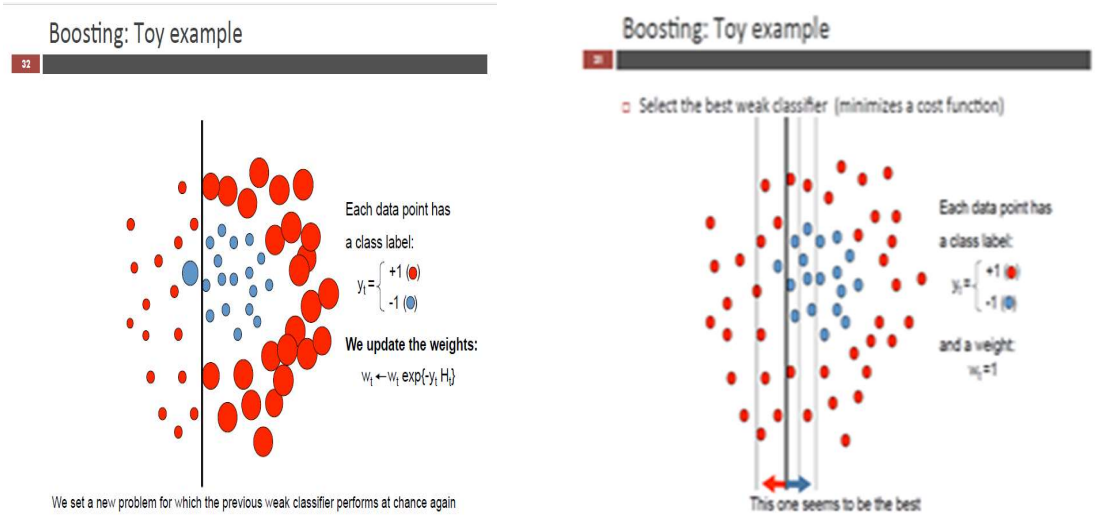


Figura 47. El primer clasificador lineal se reajustan los pesos de cada muestra clasificada incorrectamente para que el siguiente clasificador obtenga mejores resultados.

Una vez realizadas las iteraciones necesarias se obtienen una serie de clasificadores débiles que separan los datos de forma adecuada como se puede ver en la Figura 48.

La suma de los clasificadores débiles genera el clasificador fuerte como se indica a continuación: Donde h(x) es el clasificador fuerte, h_n es un clasificador débil y a_n indica el peso de cada clasificador. El clasificador lineal débil se genera analizando el histograma de valores generado al realizar la extracción de características LBP.

Con el histograma de una muestra positiva se selecciona un valor del histograma representativo de la clase, este valor se encuentra realizando varias iteraciones. Una vez tenemos un valor representativo del histograma se comparan con el resto de los histogramas de la base de datos, de forma que si un histograma no tiene este valor se considera una muestra negativa y si lo tiene se considera una muestra positiva. Este proceso interacciona tantas veces hasta conseguir un valor del histograma que genere un clasificador lineal capaz de tener una exactitud del 50%. En la Figura 49 se observa gráficamente el algoritmo de clasificación que realiza un clasificador débil.

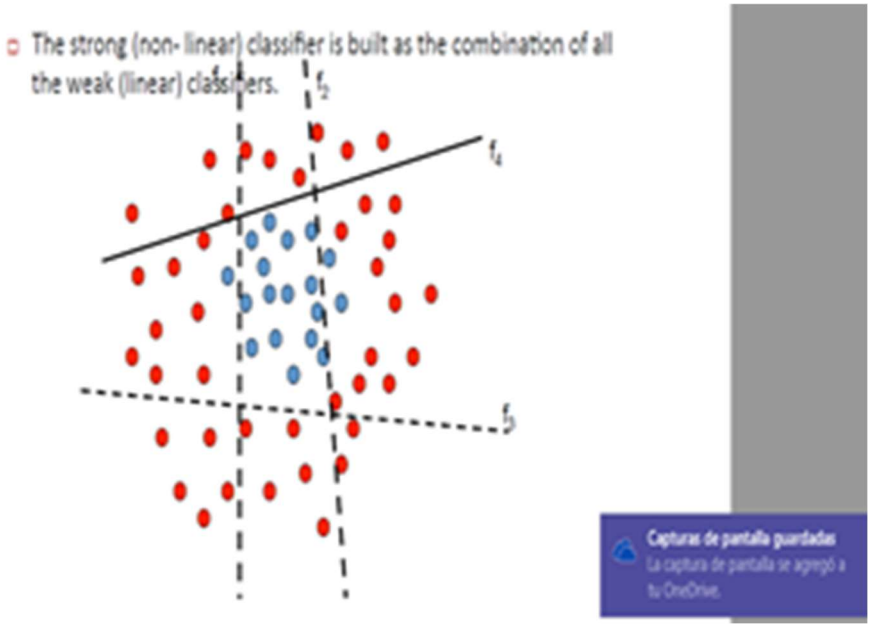


Figura 48. Clasificadores débiles finales

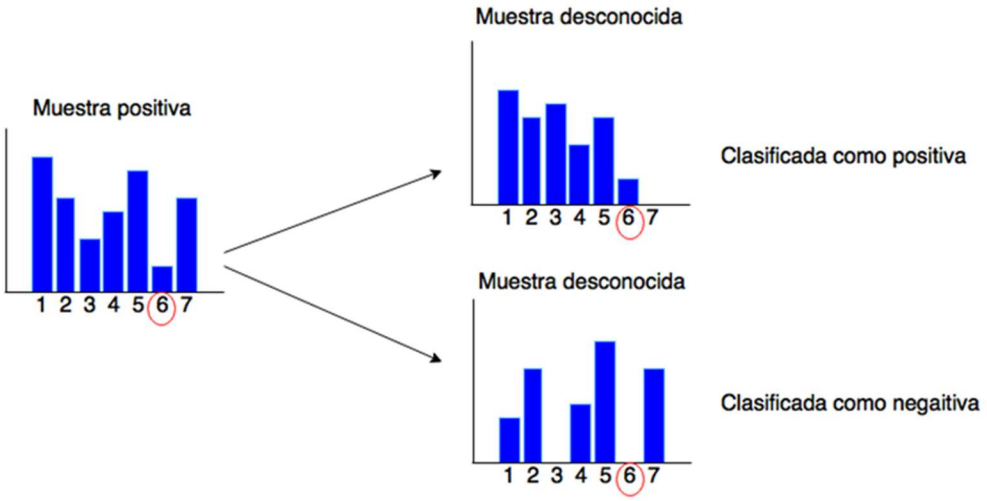


Figura 49. Clasificación mediante un clasificador débil

Una vez se consigue un clasificador débil el objetivo del clasificador Boost Cascade es conseguir que cada etapa sea mejor que la anterior, esto se consigue entrenando cada nueva

etapa con los errores cometidos en la anterior etapa de forma que las muestras falsas clasificadas como positivas son las únicas muestras negativas para la siguiente etapa.

La detección de cabezas implementada en la librería de OpenCV realiza una detección multiescala a lo largo de toda la imagen, esta detección se realiza analizando varias veces la imagen escalando el tamaño de la ventana por un factor determinado hasta que la ventana alcance el tamaño máximo posible. En nuestro caso se ha establecido un factor de escala de 1,1.

En la Figura 50 se muestra la detección multiescala para una escala 1 y 2, en la que no se detecta la cabeza.

Otro factor a tener en cuenta es que una única ventana detectada como cabeza no es suficiente para el clasificador ya que pueden existir falsos positivos con una sola detección. Este fallo se corrige teniendo en cuenta las detecciones vecinas, es decir, si se han detectado más de un número determinado de cabezas en una zona se considera el conjunto como una cabeza. En nuestro caso se ha establecido como mínimo de vecinos el valor de 5 cabezas. Esta normalización en las detecciones se muestra en la Figura 51.

Figura 50. Detección multiescala.

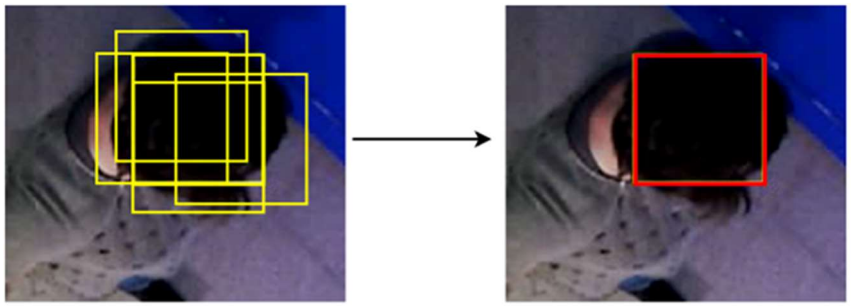


Figura 51. Normalización de la detección como mínimo de 5 cuadrados alrededor de lo detectado como cabeza.

De la Tabla 3 podemos observar que:

- La cantidad de verdaderos positivos es de 79 imágenes (VP).
- La cantidad de falsos positivos es de 22 imagen (FP).
- La cantidad de verdaderos negativos es de 101 imágenes (VN).
- La cantidad de falsos negativos es de 0 imágenes (FN).

Tabla 3. Resultados de la clasificación Boost Cascade LBP

		Predicción	
		Positivos	**Negativos**
Ground truth	**Positivos**	79	22
	Negativos	0	101

Con los valores de la Tabla 3 de la verdad se procede a calcular los diferentes parámetros definidos en las ecuaciones de la (53) a la (57). Los resultados obtenidos se muestran en la Tabla 4.

De los resultados de la Tabla 4 se puede observar que:

- El clasificador tiene un 89% de exactitud, un resultado alto.
- El clasificador tiene una precisión del 100% quiere decir que ninguna imagen negativa la ha clasificado como positiva.
- El clasificador tiene un Recall del 78%, quiere decir que no todas las imágenes positivas se han clasificado correctamente.
- El valor de F-Score es del 87%, lo que indica una relación de precisión y Recall elevada.
- El valor de Miss Rate es del 21%, esto quiere decir que la cantidad de imágenes no clasificadas correctamente es un poco notable.

Tabla 4. Resultados del test clasificador Cascade con las imágenes test.

	CASCADE
Verdaderos positivos	79
Falsos positivos	0
Verdaderos negativos	101
Falsos negativos	22
Exactitud	0,891089
Precisión	1
Recall	0,782178
F-Score	0,877778
Miss Rate	0,217821782

La segunda fase del test se ha realizado analizando tres secuencias de video. En la primera secuencia se observa una sola persona, en la segunda se observan dos personas y en la tercera secuencia se observan tres personas. Para cada secuencia se han contado cuantas cabezas había en la imagen, cuantas se han detectado correctamente y cuantas no. La detección de cabezas se realiza a nivel multiescala, de forma que, para controlar la cantidad de falsos positivos, se han puesto unos límites máximos y mínimos de tamaño de cabeza. Estos límites se establecen en la función de OpenCV que permite realizar la detección como se ve en la siguiente línea de código:

- detectMultiScale(frame_gray, heads, 1.1, 5, 0, Size(70,70), Size (100,100));

En la línea de código se puede ver que el tamaño mínimo es de 70x70 pixeles y el tamaño máximo es de 100x100 pixeles. Estos valores se han obtenido mediante prueba y error en varias ejecuciones del programa. Los resultados obtenidos en la prueba se muestran en la Tabla 5.

De los resultados que se indican en las tablas podemos observar que:

- La cantidad de falsos positivos es variable y en el caso del Grupo 2 comienza a ser notable.
- La cantidad de falsos negativos es elevada en el caso del Grupo 1 y del Grupo 2.
- El clasificador tiene una Precisión cercana al 90% esto indica que se tiene una tasa elevada de clasificaciones correctas.
- El clasificador tiene un valor de Recall muy variado, un así es un valor alto que indica que la cantidad de falsos negativos detectados depende de la situación.
- El valor de Miss Rate está por debajo del 1% en el caso del Grupo 1 y del Grupo 2, esto indica que la cantidad de imágenes verdaderas no clasificadas es baja en estos casos.
- El valor de F-score ronda el 80% y en ocasiones el 90% esto indica que el clasificador tiene una relación elevada entre la Precisión y el Recall.

Tabla 5. Resultados Boost Cascade para tres secuencias de vídeo diferentes y poder comparar.

	Resultados Boost Cascade		
	Grupo 1	**Grupo 2**	**Grupo 3**
Imágenes	200	100	50
Objetos	200	200	150
Verdaderos Positivos	187	170	137
Falsos Positivos	12	30	13
Falsos Negativos	39	49	9
Recall	0,82743	0,77625	0,93835
Precisión	0,935	0,85	0,9134
Miss Rate	0,06	0,15	0,086
F-Score	0,87793	0,81145	0,92567

El clasificador tiene un comportamiento bueno, ya que el tiempo de procesado es rápido. Se ha calculado un tiempo medio de procesado de 0,04 segundos por imagen.

En la Figura 52 se pueden observar algunos de los resultados obtenidos en la detección de cabezas en secuencias de video. En la Figura 52 (a), (c) y (e) se observan los tres casos como ha clasificado correctamente el clasificador. El primer caso, Grupo 1, es una secuencia de video en el que aparece una sola persona, el segundo caso, Grupo 2, es una secuencia de video en que aparecen dos personas y el tercer caso, Grupo 3, es una secuencia de video en que aparecen 3 personas.

Figura 52. Resultados obtenidos en la detección de cabezas en secuencias de video.

En la Figura 52 (b) se observa una detección errónea en el suelo, este tipo de falsos positivos se logran eliminar con la detección de movimiento en la imagen. En la Figura 52 (d) y (f) se observan errores de detección de falsos positivos en zonas propias de las personas, esto es debido a la similitud de características que existen, es probable que ampliando la base de datos se consigan eliminar estos errores.

Para poder comparar los dos métodos en la Tabla 6 se muestran juntas los resultados de la Tabla 2 y la Tabla 4. De los resultados se puede observar que en general todos los parámetros de SVM+HOH son mejores que los que se obtiene con CASCADE. La ventaja de usar CASCADE es su rapidez en procesar.

Tabla 6. Comparación entre los métodos SVM y Cascade.

	SVM	CASCADE
Verdaderos positivos	93	79
Falsos positivos	1	0
Verdaderos negativos	100	101
Falsos negativos	8	22
Exactitud	0,955446	0,891089
Precisión	0,989362	1
Recall	0,920792	0,782178
F-Score	0,953846	0,877778
Miss Rate	0,079207921	0,217821782

4.3.2 DETERMINACIÓN DE ÁNGULO YAW

Partimos del supuesto que la cabeza se ha encontrado mediante alguno de los métodos que se han comentado en el punto anterior. El primer paso es tomar una imagen aumentando el tamaño alrededor de la cabeza llegando hasta 240x240 pixeles. La imagen resultante se muestra en la

Figura 53. Después realizamos un *Canny* de la imagen (Figura 54), dicho algoritmo es de múltiples etapas y sirve para detectar una amplia gama de bordes en imágenes. A continuación, aplicamos una dilatación al resultado del algoritmo de *Canny* e invertimos (Figura 55). Sobre el resultado buscamos los contornos y filtramos por tamaño. Por último, aproximamos los contornos resultantes a elipses filtrando por posición y tamaño. El resultado es una elipse localizada en la cabeza que en la mayoría de los casos indica la dirección de visionado (Figura 56).

A continuación, se realiza el algoritmo de tracking y filtrado utilizado en el procedimiento 3D (capitulo 4.2) obteniendo un resultado similar al obtenido en el caso 3D.

Figura 53. Imagen original

Figura 54. Realización del algoritmo de Canny a la

Figura 53

Figura 55. A la Figura 54 se le aplica una dilatación y se invierte.

Figura 56. Elipse localizada en la cabeza que en la mayoría de los casos indica la dirección de visionado

Con posterioridad al cálculo de cada ángulo se realizan los mismos procesos que se realizaron en el 3D y el resultado es el que se muestra en la Figura 57. El error cuadrático

medio obtenido es de 15,6º que equivale a un 4,3% sobre el límite de recorrido del ángulo Yaw.

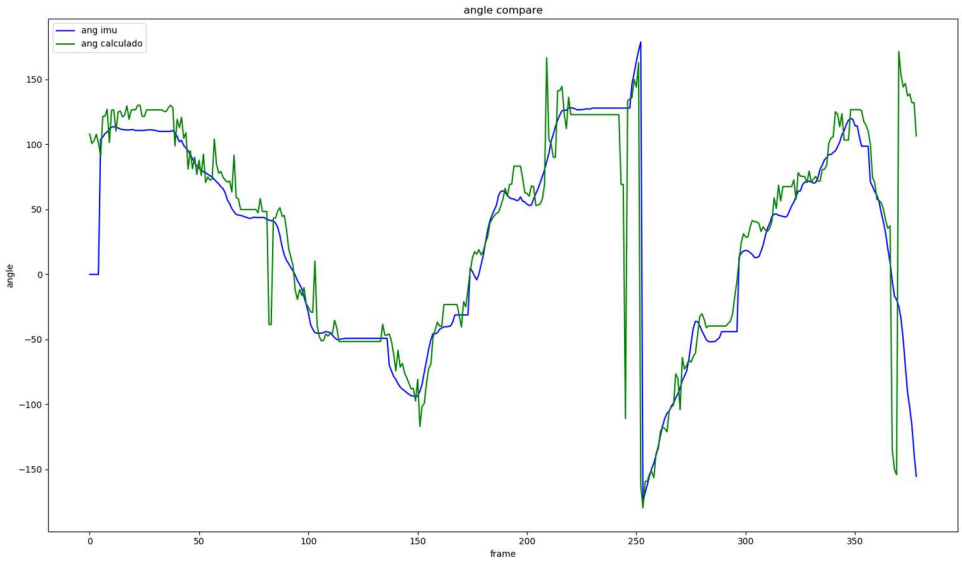


Figura 57.Comparación del ángulo calculado (verde) y ángulo medido con el IMU (azul) con el método 2D.

4.3.3 DETERMINACIÓN DE FOCO DE ATENCION 2D, DFOA Y VFOA

Partimos de la zona de análisis (Figura 1), de forma similar al método indicado en 3D; sin embargo, la zona de análisis en este caso es una matriz bidimensional, que incluye una zona con la imagen de la cámara cenital. Concretamente la zona de análisis se muestra en la Figura 58

La zona de análisis se define como,

$$ZA = M[x, y] \tag{58}$$

En dicha zona de análisis tendremos que incluir la trayectoria,

$$\text{Tr} = [\text{secuencia posiciones P(x, y)} \ \ \text{yaw}] \tag{59}$$

Figura 58: Zona de análisis con las posiciones marcadas en blanco

4.3.3.1 CALCULO DE DFOA

De las expresiones indicadas en la sección del método DFOA para cada elemento de la trayectoria se calculará con la siguiente expresión:

$$DFOA(P) = K \cdot \frac{F_{ángulo}}{d} \cdot cos\theta_N \tag{60}$$

En nuestro proceso de cálculo utilizamos una matriz real que incluye la zona de análisis, donde está incluida la imagen de la cámara, y que se utilizara para calcular las intensidades DFOA, y VFOA

$$\text{IZ(intensity zone)} = \text{Mat float (dim} = \text{dim M[x, y]l)} \tag{61}$$

Las matrices están referenciadas con el mismo eje de coordenadas respecto a las coordenadas de la zona de análisis; sin embargo, la fórmula (60) requiere que la zona de análisis se refiera a las coordenadas de la cara del sujeto en cada paso de la trayectoria. Por tanto, es necesario manipular AZ y IZ desde las coordenadas de la zona de análisis a las coordenadas del observador, realizar los cálculos y volver a transformar a las coordenadas de la zona de análisis, para realizar las acumulaciones en cada punto. Esto era necesario también realizar en la implementación 3D.

La matriz de transformación utiliza el ángulo yaw y la traslación desde el origen de la sala a la posición de la cabeza,

$$\mathrm{Mt} = [R]\ [\mathrm{T}] \tag{62}$$

Posteriormente se aplica a cada punto la expresión DFOA (60) y se realiza a continuación la transformación inversa. Nótese que en el cálculo DFOA se ha utilizado la función del ángulo, relacionado con el comportamiento de una cámara (7). El motivo es que posteriormente se confirmara la validez del método utilizando el reconocimiento de las etiquetas en las imágenes obtenidas de la cámara frontal.

4.3.3.2 CÁLCULO DE VFOA

El cálculo de la VFOA se realizará sumando las diferentes DFOA.

$$VFOA(P) = N \sum_{C} \sum_{T_r} DFOA(P) \tag{63}$$

La imagen de la Figura 59 muestra la VFOA de una trayectoria, se puede ver que hay zonas con atención muy baja o cero y otras con el círculo con atención máxima. La figura representa la magnitud de la atención en blanco en cada punto siendo la más intensa y la negra la atención nula. También está remarcada la zona con atención nula, en la que en esta trayectoria el sujeto no ha mirado en absoluto a esa zona.

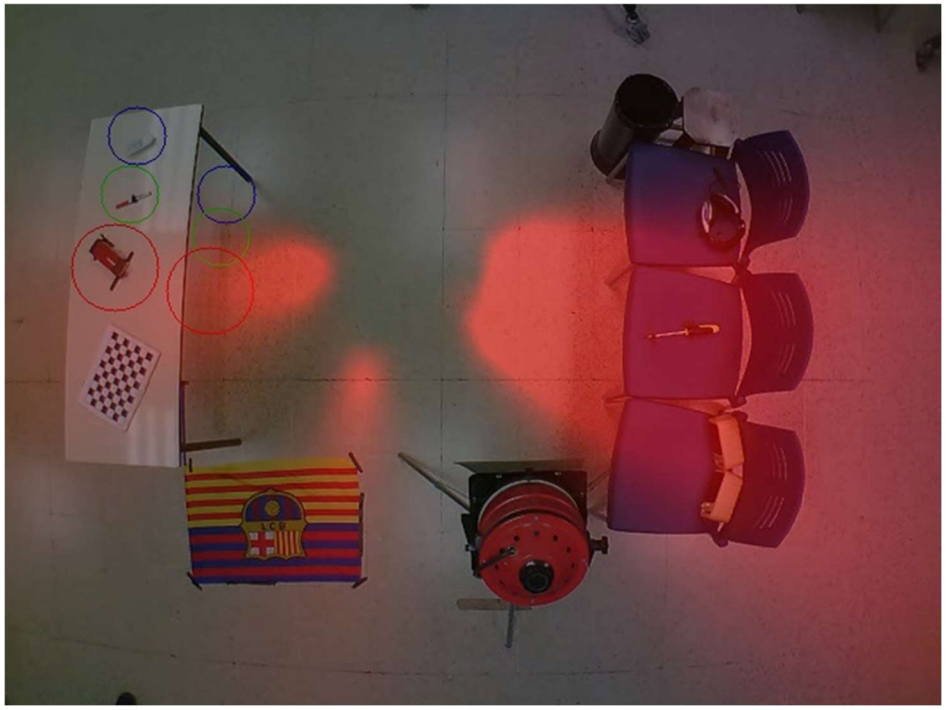


Figura 59. VFOA de una trayectoria. En rojo se observa el nivel de *VFOA*

5 VALIDACIÓN

5.1 VALIDACIÓN 3D

Al objeto de poder validar el resultado del método se han utilizado dos elementos. Por una parte, una unidad IMU, que nos ha permitido determinar los diferentes ángulos Yaw, Pitch y Roll que se utiliza para validar la determinación de la orientación de la cabeza en cada frame. Por otra parte, una cámara que tiene el mismo campo de visión que el individuo que la lleva y se utiliza para validar la atención que cada póster de color recibe de la persona. Ambos elementos estaban solidariamente ligados a la cara del sujeto mediante una estructura realizada con una impresora 3D, que encajamos en el mentón como se puede ver en la Figura 60. Los resultados obtenidos con este sistema los calificaremos de "medidos" y los hechos utilizando el método propuesto en esta tesis como "calculados".

Se ha usado diferentes localizaciones, dentro y fuera de la zona de visión de la cámara: tres colocados en las mesas (zona de visión de la cámara cenital) y tres en las paredes de la sala (fuera de la visión de la cámara cenital) para poder validar la atención en diferentes escenarios (**¡Error! No se encuentra el origen de la referencia.**). Se han dispuesto seis pósteres de colores para su fácil detección: azul-rojo, amarillo-verde, amarillo-rojo, amarillo-negro, rojo-negro y verde-negro (**¡Error! No se encuentra el origen de la referencia.**). Como se puede ver en las dos imágenes a) y b) de la Figura 60 al estar la estructura unida a la cabeza permite evaluar el ángulo Pitch, que surgirá siempre que se mire algo por debajo del nivel de los ojos de la persona. Hemos simulado esa situación con los pósteres de las mesas. Igualmente se puede validar el ángulo Yaw con todos los pósteres (**¡Error! No se encuentra el origen de la referencia.**). La comparación relativa entre la atención medida y la calculada en estos seis elementos será la regla de validación del método.

a) b)

Figura 60. Imagen mostrando cómo las personas hacen el recorrido con la cámara frontal y el IMU para hacer la validación 3D.

Figura 61. Etiquetas de verificación en paredes y mesas.

Al objeto de sincronizar las capturas se utiliza el entorno ROS (*Robot Operating System*) distribuido [87][88], capturando todas las entradas: imagen frontal, imagen cenital, imagen de profundidad (depth) cenital y la IMU de una forma sincronizada que posteriormente se procesan. Los servidores de las cámaras frontal y cenital trabajan con Linux, mientras que la IMU es con Android. Las especificaciones de los equipos son:

1. Cámara cenital: intel D435 a 6fps de 640x480 [89].
2. Cámara frontal: web cam a 30 fps de 1280x 720.
3. IMU: generación de cuaterniones [90][91], acelerómetro icm20690 y giroscopio icm20690 [92].

5.1.1 VALIDACIÓN DE LA POSICIÓN DE LA CABEZA

La validación de la posición de la cabeza se realiza a través de las imágenes adquiridas y su comparación de la posición estimada con la posición real (Figura 62).

Figura 62. El círculo rojo muestra el centro de la posición de la cabeza de la persona siendo calculada

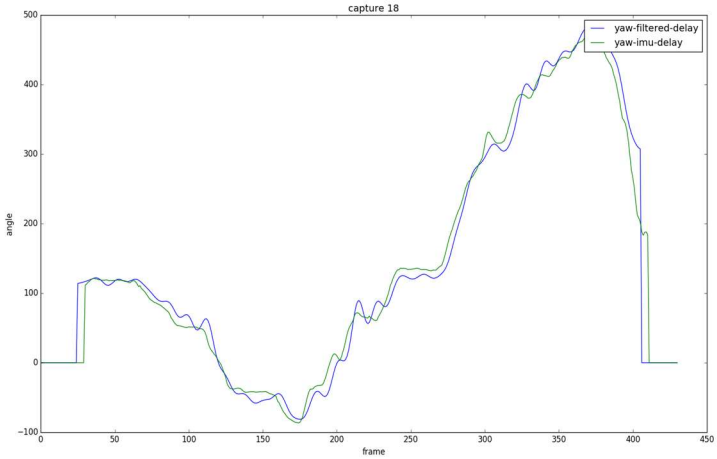
5.1.2 VALIDACIÓN DE ÁNGULOS YAW, PITCH Y ROLL

La validación se realiza mediante el uso de una unidad inercial IMU (icm20690), del que se obtienen cuaterniones y una marca temporal (*timestamp*), a un ancho de banda de 100Hz. La marca temporal es una secuencia de caracteres que denotan el tiempo en segundos a nanosegundos (*Unix Time Stam*) en las que se han tomado los cuaterniones. Esta información permite la comparación entre dos registros diferentes y el seguimiento de avances en el tiempo fácilmente. De los cuaterniones se obtienen los ángulos de Euler (Yaw, Pitch y Roll). Como se ha razonado anteriormente para nosotros los interesantes son el Yaw y el Pitch.

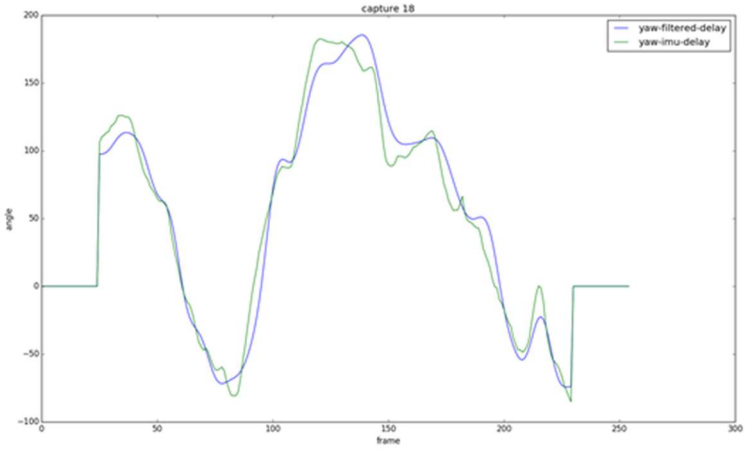
Por otra parte, la captura de la cámara cenital (D435) obtienen las imágenes de profundidad y la imagen en color. También dispone de un marcador temporal como la webcam (imagen frontal).

Utilizando las imágenes de la cámara cenital se obtiene por los procedimientos indicados anteriormente los ángulos (Yaw, Pitch) asociados a una imagen. Y por lo tanto disponemos de los ángulos calculados para cada tiempo de captura de la imagen. A continuación, buscamos en la secuencia de los cuaterniones contenidos en la IMU cuál es el más próximo al marcador temporal de la imagen de la cámara, esto es, donde el error de tiempos será despreciable. Decimos esto porque la cámara captura 6 Hz y el IMU trabaja a 100 Hz, con lo que el error de tiempo de captura será de un 6% (6/100). Consideramos que la cabeza humana prácticamente no se mueve entre frames, por lo tanto, el movimiento en un 6% del tiempo de frame lo consideramos despreciable.

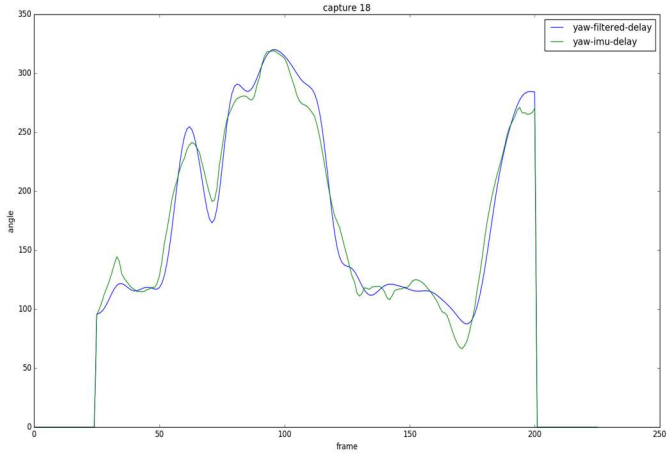
En la *Figura 63* mostramos algunas comparaciones de trayectorias donde se muestra el ángulo Yaw medido (verde) y el ángulo Yaw calculado (azul) y se indica de qué captura se trataba. Como se puede ver en la figura las trayectorias a), b) y c) el ángulo Yaw es bien diferente entre ellas porque cada persona habrá realizado una trayectoria diferente fijándose de forma diferente en los objetos de análisis y dedicando un tiempo diferente a cada uno de ellos. No siempre todos los objetos son observados. En cada caso se ha medido el error medio



a)



b)



c)

Figura 63. Comparación del ángulo Yaw calculado (azul) y medido (verde).

obtenido para el ángulo Yaw que se muestra en la Tabla 7. El error medio, en valor absoluto que se obtiene lo consideramos pequeño.

La misma operación se ha hecho con el ángulo Pitch. Mostramos tres casos como para el ángulo Yaw, ya que el resto de los resultados son similares a los que mostramos. En la Figura 64 se muestra tres trayectorias comparando para cada una de ellas el ángulo Pitch medido (verde) y el calculado (azul). En la Tabla 8 se muestra el error medio obtenido para el ángulo Pitch. Los resultados obtenidos tienen menor error para este ángulo que para el ángulo Yaw.

Aquí se ha mostrado los resultados de tres trayectorias, pero se han tenido en cuenta cincuenta trayectorias. En este caso el error medio en valor absoluto para el ángulo Yaw es 11,5º, para el ángulo Pitch es 4,4º.

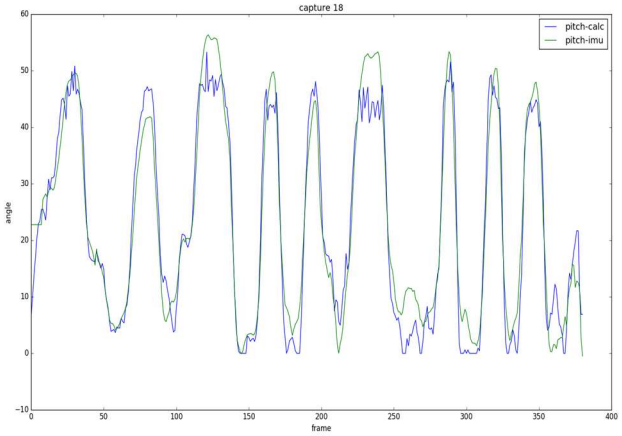
El ángulo de Roll medido con la unidad IMU tiene una media cuadrática de 3,3º; por lo que considerando el valor de 0º para todas las posiciones de nuestra trayectoria nuestro error será de 3,3º , inferior a cualquiera de los otros ángulos .

Tabla 7. Error medio en valor absoluto del ángulo Yaw calculado y medido de las *Figura 63*.

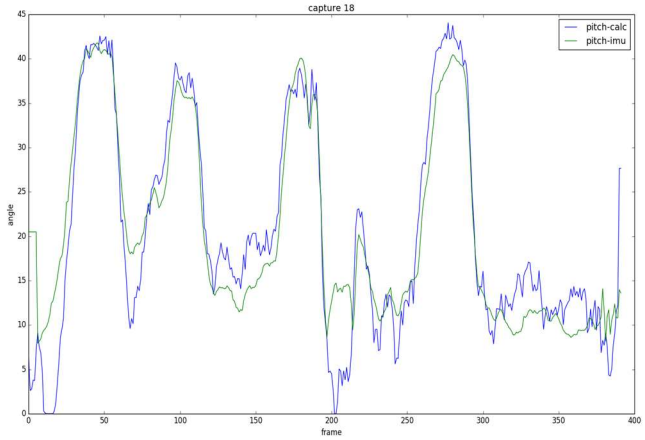
	Figura 63 a)	*Figura 63* b)	*Figura 63* c)
Error medio, en valor absoluto, del ángulo Yaw calculado y medido	15º	9,2º	8,1º

Tabla 8. Error medio en valor absoluto del ángulo Pitch calculado y medido de la Figura 64

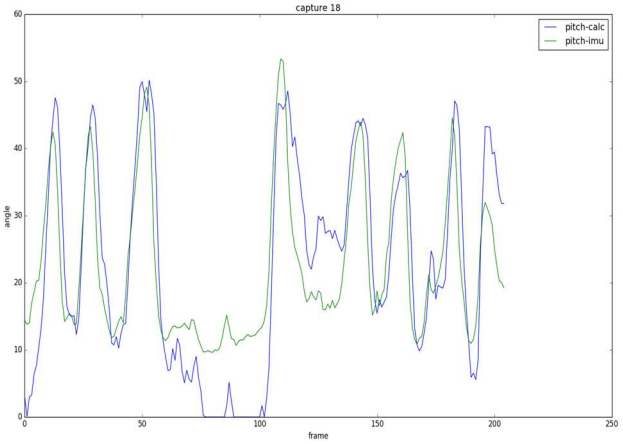
	Figura 64**a)**	Figura 64 **b)**	Figura 64 **c)**
Error medio, en valor absoluto, del ángulo Pitch calculado y medido	3,7º	3,5º	7,2º



a)



b)



c)

Figura 64. Comparación del ángulo Pitch calculado (azul) y medido (verde).

5.1.3 MÉTODO DE VALIDACIÓN VFOA

La validación final se realiza comparando los valores asociados a los objetivos (pósteres de colores, **¡Error! No se encuentra el origen de la referencia.**). La validación de DFOA, aun siendo aproximada, permite ver si hay errores importantes en algún elemento de una trayectoria. En la Figura 65 se puede observar como los bordes de la imagen coinciden con el límite de la zona DFOA en la nube de puntos. La DFOA permite comprender de una forma visual la bondad del método. Sin embargo, vamos a pasar a dar la validación de la VFOA.

Utilizando la cámara frontal (Figura 60) se detecta los diferentes pósteres de doble color. El método utilizado para detectar dichas etiquetas consiste en la secuencia siguiente:

1. Detección de los colores de los bordes (utilizando comparación en el espacio de color HSV) y generación de las máscaras con los valores comprendidos en el rango.
2. Generación de contornos asociados a las máscaras generadas en el paso 1.
3. Filtrado de la superficie de los contornos anteriores por rango.
4. Búsqueda del color interior, dentro de los elementos filtrados en el punto 3.
5. Filtrado por superficie del segundo color.

La detección de etiquetas se realiza frame a frame del video asociado a la cámara frontal. Como ya hemos mencionado en el apartado 5.1.2, el tiempo de frame de la cámara frontal, no coincide con el tiempo de frame de la cámara de captura de las trayectorias, la cámara cenital. Y por tanto el método de comparación tendrá que tener esto presente y solucionarlo.

Como método de análisis de la atención fijada en cada etiqueta, contamos el número de pixeles, que se detecta en cada trayectoria por la cámara frontal, que haciendo un símil con un ojo sería equivalente al impacto que deja una etiqueta en los sensores internos del mismo. El número de pixeles total de cada color de cada trayectoria es proporcional a la atención de una etiqueta determinada en dicha trayectoria (ver el capítulo 3 con la descripción del método):

$$NPixel_i = A \sum_{P_{label_i}} VFOA(P) \tag{64}$$

Donde P son puntos que pertenecen al *label* asociado.

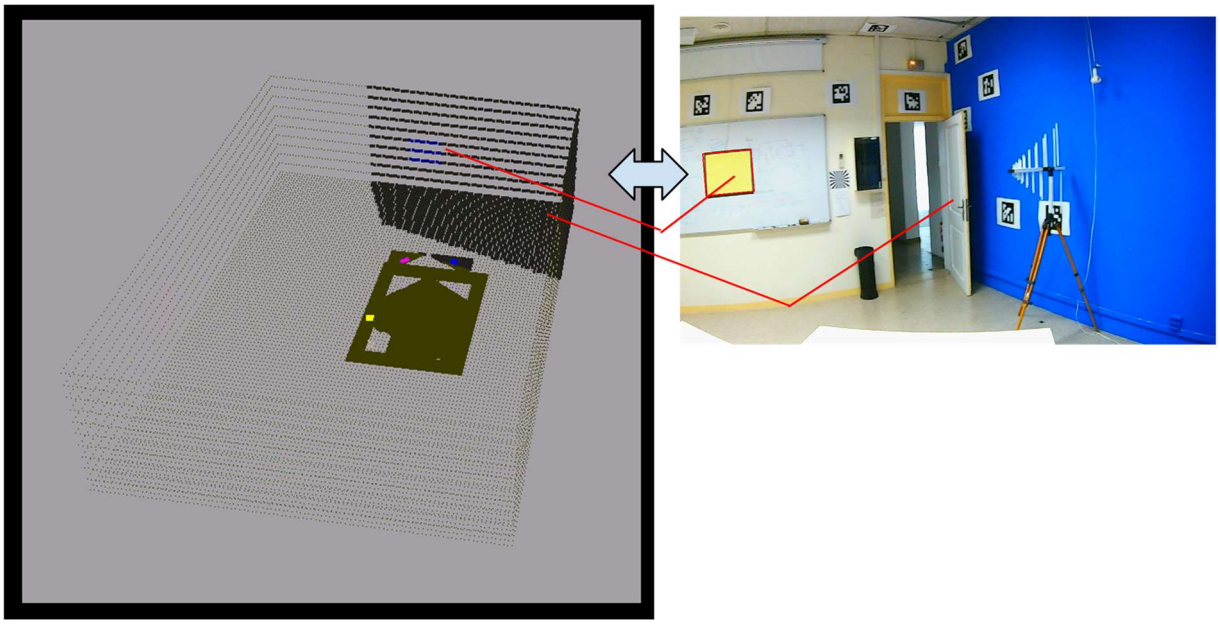


Figura 65. Comparación de la zona DFOA activa (izquierda) y la imagen equivalente (derecha).

Para poder realizar la comparación del método y como no conocemos la relación de proporcionalidad, solo podemos obtener una relación entre los pósteres. Por lo que en las comparaciones fijaremos un objetivo como referencia. La relación, entre $NPixel_i$ / $NPixel_{Ref}$ nos indica el porcentaje de atención a un objetivo medido. Y la relación entre $\sum_{P_{label_i}} VFOA(P)$ /$\sum_{P_{label_{referencia}}} VFOA(P)$ nos dará la comparación entre objetivos calculados. Este método, no limita la usabilidad del VFOA, puesto que al final del recorrido lo que deseamos es saber qué póster ha sido más visto y en qué relación.

El error en cada trayectoria viene dando:

$$Error = \left| \frac{Nlabel_i}{Nlabel_{referencia}} \right| - \left| \frac{\sum_{P_{label_i}} VFOA(P)}{\sum_{P_{label_{referencia}}} VFOA(P)} \right| \quad (65)$$

Para cada *label*,.se cuentan los puntos captados por la cámara frontal y se compara con los puntos de un *label* de referencia. Y se verifica que la relación de VFOA correspondiente a la misma *label* con la VFOA del *label* de referencia.

La razón de realizar estas divisiones es meramente por un tema dimensional; los puntos de la imagen de la cámara frontal tendrán dimensión de PUNTO mientras que VFOA tiene dimensión de TIEMPO/DISTANCIA^2, con lo que la única forma de compararlas es a través de la relación que existen entre ellas.

El resultado de esta comparación teniendo en cuenta cincuenta trayectorias es que el error medio del VFOA es del 16% y el error máximo es del 32%.

5.2 VALIDACIÓN 2D

En el proceso de validación de los métodos 2D se han utilizado elementos subjetivos. Estos métodos han estado concentrados en la validación de la detección y posicionamiento de la cabeza y la validación del ángulo, de esta forma queda validado el método [1]. Adicionalmente se ha validado también el método de cálculo de los tiempos de atención. Por una parte, en la detección de las cabezas el método de validación fue la inspección visual del resultado obtenido con el método. En cada frame el resultado de la detección se enmarca en la imagen de esta forma se ha revisado cada imagen visualmente y se ha comprobado su éxito o error. Con respecto a la validación del ángulo una validación visual no era posible y se partió de las cabezas detectadas en el método 3D donde cada individuo tenía conectado una IMU solidaria con la cabeza, con lo que la validación del ángulo se realizó en base a la comparación con los datos de la IMU.

Por último, la validación de la VFOA calculada se realiza por comparación entre la opinión del individuo y del que hemos llamado inspector, ambos anotaban el recorrido y donde prestaba atención la persona y durante cuánto tiempo mientras se estaba grabando.

5.2.1 SEGUIMIENTO

El análisis temporal del seguimiento permite también calcular correctamente la dirección de la cabeza (Figura 66): calculamos la dirección en la detección inicial basada en el punto de entrada en el campo de la cámara y el movimiento de la cabeza. En los *frames* subsiguientes, el cambio en la dirección de *frame* a *frame* se va calculando y está restringido porque no se permiten giros de 180° de *frame* a *frame*.

Figura 66. Dirección y máscara de la elipse de la cabeza.

5.2.2 MÉTODO DE VALIDACIÓN DE LA ATENCIÓN EN OBJETIVOS, (VALIDACIÓN VFOA)

El método de valoración en 2D es un método subjetivo, que requiere de una encuesta a las personas que han realizado el test y la presencia de un inspector que analiza a que objetos ha mostrado atención la persona del test. De forma que el sujeto indica después del test con qué porcentaje ha visualizado cada uno de los objetos, de igual forma el inspector también indica con qué porcentaje ha visualizado cada uno de los objetos.

Estos resultados se llevaron a cabo por un alumno TFG [92]. En donde se realizaron dos pruebas con diferente escenario, en la primera prueba participaron 11 individuos (Figura 67 a)) y en la segunda 20 (Figura 67b)). La zona de análisis tiene forma de U con la entrada

para que la gente se mueva y mire aquello que le atraiga. Ambas zonas están en el área de captación de la cámara, en la a) se extiende hasta los límites y en la b) se estrecha para evitar las deformaciones de los extremos. En ambos escenarios hay diez objetos. Los resultados se analizaron por objetos y por zonas de forma que se consideró zona de objetos a la derecha como Derecha y zona de objetos a la izquierda como Izquierda.

En la Tabla 9 se muestra la suma de los errores en valor absoluto con respecto a cada objeto, y con respecto a las zonas derecha e izquierda para los dos escenarios. En primer lugar, se obtiene menor error utilizando el método 3D. A la luz de los resultados, entendemos que el resultado 2D es poco preciso con objetos de dimensión pequeña y próximos entre sí que es como estaban colocados (Figura 67). Sin embargo, cuando no queremos precisión y tratamos con objetos grandes o zonas grandes, el método 2D da errores aceptablemente pequeños.

a) b)

Figura 67. Dos zonas de análisis donde la gente se desplaza y observa lo que le llama la atención

Tabla 9. Errores sumados en valor absoluto con respecto a cada objeto, y con respecto a las zonas.

MÉTODO	Prueba 1 OBJETOS	Prueba 1 Zonas	Prueba 2 OBJETOS	Prueba 2 Zonas
Método 2D	39%	8,2%	32%	6,5%
Método 3D	27%	5,3%	27%	4,6%

5.2.3 MÉTODO DE VALIDACIÓN DE LOS TIEMPOS DE ATENCIÓN

Ya comentamos que el método propuesto puede evaluar también los tiempos que un sujeto está mirando a un determinado punto. Pudiendo así simular la colocación de cámaras que detecten caras y tiempos de atención en cada punto de la zona de análisis.

La forma de validación la hicimos mediante sujetos instruidos para hacer recorridos con un trayecto y atención pactados de antemano. Existieron dos tipos de trayectorias: lineales y circulares. En las trayectorias lineales (marcadas en el piso por tiras de cinta de diferentes colores) (Figura 68 a)), se les indica a los individuos que miren un solo póster en la trayectoria completa (Figura 69). Hay cuatro franjas por las que se puede caminar, en dos direcciones y cuatro carteles, lo que da como resultado 32 combinaciones. Para las trayectorias circulares, los individuos miran consecutivamente el cartel más cercano (uno diferente en cada cuadrante de la trayectoria, Figura 69).

Cada trayectoria dura 12,5 segundos o 250 cuadros (grabados a 20 Hz). Se han registrado individuos de diferentes características: hombres y mujeres, con pelo y sin pelo, altos y bajos, con sombrero y sin sombrero.

Para cada grabación, la posición y la dirección de la cabeza en cada cuadro se anota manualmente, para crear datos reales de las trayectorias. Esto permitirá validar el método mediante la comparación de los tiempos calculados automáticamente utilizando el algoritmo propuesto y los datos reales. Como nuestro objetivo es demostrar la validez del método para medir los tiempos, hemos optado por la anotación manual en lugar de utilizar el algoritmo de seguimiento [1]. Para cada trayectoria registrada, las anotaciones de la trayectoria manual se han utilizado para calcular los ángulos de visión desde cada punto de la trayectoria hasta los carteles correspondientes.

En la Figura 70 se muestra el análisis temporal, tiempo de atención, visualización y permanencia para los resultados de las trayectorias circular. Para cada trayectoria registrada, las anotaciones de la trayectoria manual se han utilizado para calcular los ángulos de visión desde cada punto de la trayectoria hasta los carteles correspondientes. Los ángulos se han

a) b)

Figura 68. Ejemplo de trayectorias: a) lineal; b) curva. Los vectores en cada punto de la trayectoria indican la dirección de la cabeza.

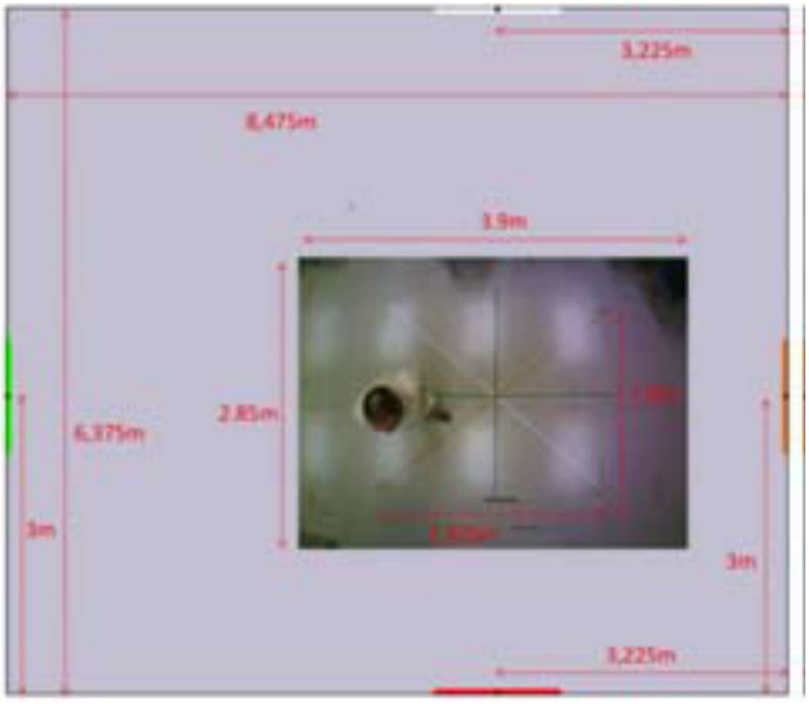


Figura 69. Setup experimental con los cuatro posters (blanco, naranja, roja, verde) en las paredes para ser analizados.

cuantificado utilizando intervalos de 5 grados, y se ha calculado un histograma de los ángulos.

Los tiempos de permanencia, visualización y atención se han calculado usando

$$T^a = \sum_i T_i^a, \qquad T^{iv} = \sum_i T_i^{iv} \tag{66}$$

donde T_i^a es el tiempo de atención (A*ttention*) y T_i^{iv} es el tiempo de visualización (*In-view*). El tiempo de permanencia se basa en el número de cuadros en los que el individuo se detecta por completo. El tiempo de visualización se calcula como el número de frames consolidados a 45º. Un cuadro o frame se considera como consolidado en un ángulo dado α si durante los siguientes 20 frames (1s) el ángulo de visión |θ - ϕ | permanece igual o menor que α (Figura 13a). Consideramos los casos en los que |θ - φ | es lo suficientemente pequeño para que el póster esté dentro del cono de visión de la persona teniendo una atención completa, esto es, cuando se cumple |θ - φ | ≤ 25º. Por eso se dice que el tiempo de atención se calcula utilizando un ángulo de consolidación de 25º.

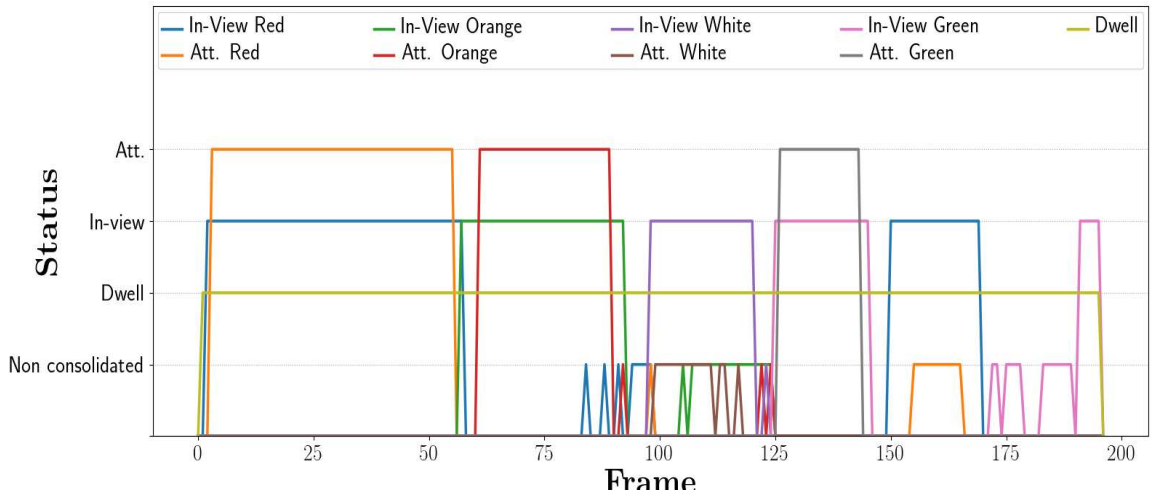


Figura 70. Análisis temporal de la trayectoria circular.

Para validar el método, se le ha pedido a un observador independiente que visualice los videos e indique el tiempo en frames (denominado tiempo del observador,) en el que estima que el individuo sometido a prueba está mirando un objetivo específico. Los resultados se resumen en la Figura 70 y la Tabla 10. Para cada trayectoria y cada objetivo, se muestran los tiempos de visualización calculados ({Dwell, T_i^d}, {In-view, T_i^{iv}}, {Atención, T_i^a}). Los tiempos se indican en fotogramas (fr) y su equivalente en segundos (s). En la trayectoria lineal, la persona de prueba mira el objetivo rojo durante toda la grabación. En la trayectoria circular, la persona mira alternativamente y consecutivamente a los cuatro objetivos.

Tabla 10. Tiempos de visualización en un recorrido lineal y circular.

Trayectoria	Patrón	Valores		
		T_i^d	T_i^{iv}	T_i^a
Lineal	Rojo	70 fr 3,50 s	67 fr 3,35 s	63fr 3,15s
	Naranja	70 fr 3,50 s	0 fr 0,00 s	0 fr 0,00 s
	Blanco	70 fr 3,50 s	0 fr 0,00 s	0 fr 0,00 s
	Verde	70 fr 3,50 s	0 fr 0,00 s	0 fr 0,00 s
Circular	Rojo	195 fr 9,75 s	76 fr 3,80 s	53 fr 2,65 s
	Naranja	195 fr 9,75 s	36 fr 1,80 s	29 fr 1,45 s
	Blanco	195 fr 9,75 s	23 fr 1,15 s	0 fr 0,00 s
	Verde	195 fr 9,75 s	26 fr 1,30 s	20 fr 1,00 s

Los errores encontrados difieren para cada uno de los tiempos Dwell, T_i^d es 0 debido a que la detección de presencia no detecta errores. Con respecto a los tiempos In-view, T_i^{iv} y Atención, T_i^a los errores son 4,6% y 5,4% respectivamente.

Para realizar esta validación se realizaron 21 trayectorias rectilíneas y 20 trayectorias circulares [1].

6 CONCLUSIONES

El objetivo principal de la tesis era encontrar un método que permitiese evaluar la atención de las personas que transitan por una determinada zona, utilizando un número reducido de cámaras. El método debía de poder tener unos resultados similares a los métodos anteriores, que utilizaban una cámara localizada en cada objeto de interés.

En esos métodos se mide el tiempo en el que la cara mira a esa cámara. Pero no se tiene en cuenta si la cara está cerca o lejos del objeto. Por ello, también era un objetivo de la tesis buscar una métrica de la atención más precisa que incluyese el factor distancia.

En el trabajo presentado se ha analizado el comportamiento del ojo como sensor y descrito un método que cuantifica la cantidad de sensaciones que se han podido formar en el interior de este. Con esa base se ha presentado una expresión que permite calcular y visualizar, en una herramienta estándar de representación de superficies (PCL), las magnitudes de atención que se han tenido sobre las mismas, es decir la función VFOA. Así mismo el método permite encontrar las magnitudes de tiempo de la métrica anterior, pero con la ventaja de que se obtienen para cualquier punto de la zona de análisis. En los métodos anteriores solo es posible calcularlo en los objetos que disponen de cámara frontal.

Por otra parte, queríamos realizar implementaciones de bajo coste; es decir comercializables. Este objetivo está también cumplido tal como se explica en el capítulo de implementación 2D. Sin embargo, es muy posible que los algoritmos de implementación que hemos utilizado en esta primera demostración del método vayan evolucionando en función de la existencia de mejoras en las librerías de procesado de imagen. Lo que hemos obtenidos con los dos métodos estudiados es que en general todos los parámetros de SVM+HOH son mejores que los que se obtiene con CASCADE. La ventaja de usar CASCADE es su rapidez en procesar.

Dentro de las implementaciones, hemos justificado que solo los métodos que utilizan la profundidad pueden ser precisos. Los métodos que utilizan cámaras RGB sin profundidad solo pueden tener resultados aproximados en áreas grandes. Que, si bien es posible que utilicemos de una manera económica, solo pueden dar resultados aproximados.

En la validación 3D teniendo en cuenta cincuenta trayectorias el error medio del VFOA es del 16% y el error máximo es del 32%. El error medio en valor absoluto para el ángulo Yaw es 11,5º y para el ángulo Pitch es 4,4º.

7 LÍNEAS FUTURAS

Este proyecto es un doctorado industrial que tiene como objetivo generar las bases científicas y tecnológicas para desarrollar un producto comercial. Con esta tesis no se ha acabado el proyecto. Los trabajos futuros serían:

- **Sincronización de una red de cámaras cenitales**

El producto final debe contemplar poder cubrir zonas amplias. Para ello necesitamos aumentar el número de cámaras cenitales y formar una red de cámaras que cubra la zona. Ello nos obligara a cambiar el algoritmo de detección de personas y cabezas utilizando varias cámaras, y a sincronizar la información de cada una de ellas.

Adicionalmente tendremos que generar un hardware adecuado y compacto que sea instalable por cualquier instalador del mercado, y un software que pueda entregar datos estadísticos en tiempo real para el cliente.

Este objetivo es imprescindible para su comercialización y requerirá tanto de ajustes en los algoritmos como de la generación de hardware que permita la centralización y el procesado simultaneo de resultados.

- **Aumento de la base de datos de personas**

Hasta ahora estamos trabajando con pocas personas para realizar las capturas y los algoritmos deben de actualizarse con bases de datos de miles de personas para que los resultados sean comercialmente fiables.

- **Caracterización de los individuos**

En los trabajos presentados en esta tesis, solo se analiza la atención de la persona. Sin embargo, es de alto interés comercial no solo saber que una persona ha mirado cierto producto o anuncio; sino también qué características tiene esa persona. Es decir, si se trata de un hombre o una mujer, el tipo social, el rango de edad, etc...

Esto permitiría cruzar características de que cosas les gusta o preocupa a una cierta clase de personas y ofrecer esta información comercial a los clientes. El paso para proporcionar estos datos es fácilmente enlazable con el método propuesto y seguro que tendremos trabajos futuros en esta línea.

8 REFERENCIAS BIBLIOGRÁFICAS

[1] M. Lopez-Palma, J. R. Morros, J. Gago, and M. Corbalan, “Oriented Trajectories as a Method for Audience Measurement,” in *IEEE International Symposium on Industrial Electronics*, 2018, vol. 2018-June.

[2] BOE, “BOE.es - Documento BOE-A-2018-16673,” *«BOE» núm. 294, de 6 de diciembre de 2018, páginas 119788 a 119857 (70 págs.) Jefatura del Estado*, 2018. [Online]. Available: https://www.boe.es/eli/es/lo/2018/12/05/3. [Accessed: 28-Aug-2020].

[3] C. Papageorgiou and T. Poggio, “A Trainable System for Object Detection,” *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, Jun. 2000.

[4] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, “Survey of pedestrian detection for advanced driver assistance systems,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.

[5] A. Broggi, A. Fascioli, I. Fedriga, A. Tibaldi, and M. Del Rose, “Stereo-based preprocessing for human shape localization in unstructured environments,” in *IEEE Intelligent Vehicles Symposium, Proceedings*, 2003.

[6] A. Mohan, C. Papageorgiou, and T. Poggio, “Example-based object detection in images by components,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001.

[7] A. Datta, M. Shah, and N. Da Vitoria Lobo, “Person-on-person violence detection in video data,” *Proc. - Int. Conf. Pattern Recognit.*, 2002.

[8] P. Viola, M. J. Jones, and D. Snow, “Detecting pedestrians using patterns of motion and appearance,” *Int. J. Comput. Vis.*, 2005.

[9] N. Dalal, B. Triggs, and C. Schmid, “Human detection using oriented histograms of flow and appearance,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2006.

[10] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, “Robust tracking-by-detection using a detector confidence particle filter,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2009.

[11] H. Cho, Y. W. Seo, B. V. K. V. Kumar, and R. R. Rajkumar, “A multi-sensor fusion system for moving object detection and tracking in urban driving environments,” in *Proceedings - IEEE International Conference on Robotics and Automation*, 2014.

[12] O. Ozturk, Toshihiko Yamasaki, and Kiyoharu Aizawa, “Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis,” in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, 2009, pp. 1020–1027.

[13] M. Andriluka, S. Roth, and B. Schiele, “People-tracking-by-detection and people-detection-by-tracking,” in *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008.

[14] C. C. Chen, H. H. Lin, and O. T. C. Chen, “Tracking and counting people in visual surveillance systems,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2011.

[15] L. Kratz and K. Nishino, “Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.

[16] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian Detection: An Evaluation of the State of the Art,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.

[17] S. Zangenehpour, L. F. Miranda-Moreno, and N. Saunier, “Automated classification based on video data at intersections with heavy pedestrian and bicycle traffic:

Methodology and application," *Transp. Res. Part C Emerg. Technol.*, vol. 56, pp. 161–176, 2015.

[18] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Comput. Vis. Image Underst.*, 2007.

[19] F. A. Cheikh, J. Y. Hardeberg, D. Gouton, PierLefloch, Re, and R. Picot-Clemente, "Real-time people counting system using a single video camera," *Real-Time Image Process. 2008*, 2008.

[20] T. Y. Chen, C. H. Chen, D. J. Wang, and Y. L. Kuo, "A people counting system based on face-detection," in *Proceedings - 4th International Conference on Genetic and Evolutionary Computing, ICGEC 2010*, 2010.

[21] A. Virgona, A. Alempijevic, and T. Vidal-Calleja, "Socially constrained tracking in crowded environments using shoulder pose estimates," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2018, pp. 4555–4562.

[22] J. Clement, J. Aastrup, and S. Charlotte Forsberg, "Decisive visual saliency and consumers′ in-store decisions," *J. Retail. Consum. Serv.*, vol. 22, pp. 187–194, Jan. 2015.

[23] S. C. Kuo, C. J. Lin, and C. C. Peng, "Using Adaboost Method for Face Detection and Pedestrian-Flow Evaluation of Digital Signage," in *2014 International Symposium on Computer, Consumer and Control*, 2014, pp. 90–93.

[24] M. Farenzena, L. Bazzani, V. Murino, and M. Cristani, *No Title*, vol. 5716 LNCS. Springer, Berlin, Heidelberg, 2009, pp. 481–489.

[25] T. Wästlund, Erik and Shams, Poja and Otterbring, "Unsold is unseen ... or is it? Examining the role of peripheral vision in the consumer choice process using eye-tracking methodology," *Appetite*, vol. 120, pp. 49--56, 2018.

[26] V. Drouard, R. Horaud, A. Deleforge, S. Eye Ba, and G. Evangelidis, "Robust Head-

Pose Estimation Based on Partially-Latent Mixture of Linear Regressions."

[27] F. Kuhnke and J. Ostermann, "Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, vol. 2019-Octob, pp. 10163–10172.

[28] R. Ravnik and F. Solina, "Audience measurement of digital signage: Quantitative study in real-world environment using computer vision," *Interact. Comput.*, vol. 25, no. 3, pp. 218–228, 2013.

[29] R. Ravnik, F. Solina, and V. Zabkar, "Modelling In-Store Consumer Behaviour Using Machine Learning and Digital Signage Audience Measurement Data," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8811, pp. 123–133, 2014.

[30] R. Ravnik and F. Solina, "Interactive and audience adaptive digital signage using real-time computer vision," *Int. J. Adv. Robot. Syst.*, vol. 10, no. 2, p. 107, Feb. 2013.

[31] G. E. S. Battiato, A. Cavallaro, and C. Distante, "Special issue on 'Video analytics for audience measurement in retail and digital signage,'" *Pattern Recognition Letters*, vol. 81. pp. 1–2, 2016.

[32] J. Zieren, N. Unger, and S. Akyol, "Hands Tracking from Frontal View for Vision-Based Gesture Recognition," 2007.

[33] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 300–311, Jun. 2010.

[34] J. Clement, J. Aastrup, and S. Charlotte Forsberg, "Decisive visual saliency and consumers' in-store decisions," *J. Retail. Consum. Serv.*, vol. 22, pp. 187–194, 2015.

[35] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans.*

Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 185–207, Jan. 2013.

[36] R. Stiefelhagen, "Tracking focus of attention in meetings," in *Proceedings - 4th IEEE International Conference on Multimodal Interfaces, ICMI 2002*, 2002, pp. 273–280.

[37] M. Voit and R. Stiefelhagen, "Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios," in *ICMI'08: Proceedings of the 10th International Conference on Multimodal Interfaces*, 2008, pp. 173–180.

[38] S. O. Ba and J. M. Odobez, "Multiperson visual focus of attention from head pose and meeting contextual cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 101–116, Jan. 2011.

[39] S. Duffner and C. Garcia, "Visual Focus of Attention Estimation with Unsupervised Incremental Learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 12, pp. 2264–2272, Dec. 2016.

[40] B. Masse, S. Ba, and R. Horaud, "Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2711–2724, 2018.

[41] M. Quintana, J. M. Menendez, F. Alvarez, and J. P. Lopez, "Improving retail efficiency through sensing technologies: A survey," *Pattern Recognit. Lett.*, vol. 81, pp. 3–10, Oct. 2016.

[42] M. Ariz, A. Villanueva, and R. Cabeza, "Robust and accurate 2D-tracking-based 3D positioning method: Application to head pose estimation," *Comput. Vis. Image Underst.*, vol. 180, pp. 13–22, 2019.

[43] E. N. Arcoverde Neto *et al.*, "Enhanced real-time head pose estimation system for mobile device," *Integr. Comput. Aided. Eng.*, vol. 21, no. 3, pp. 281–293, Apr. 2014.

[44] C. Tang and Q. Chen, "Zenithal people counting using histogram of oriented gradients," in *2012 5th International Congress on Image and Signal Processing, CISP*

2012, 2012, pp. 946–951.

[45] L. Del Pizzo, P. Foggia, A. Greco, G. Percannella, and M. Vento, “Counting people by RGB or depth overhead cameras,” *Pattern Recognit. Lett.*, vol. 81, pp. 41–50, 2016.

[46] M. Castrillón-Santana, J. Lorenzo-Navarro, and D. Hernández-Sosa, “Conteo de personas con un sensor RGBD comercial,” vol. 11, no. 3, pp. 348–357, Jul. 2014.

[47] M. Rauter, “Reliable human detection and tracking in top-view depth images,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 529–534.

[48] M. Sturari *et al.*, “Robust and affordable retail customer profiling by vision and radio beacon sensor fusion,” *Pattern Recognit. Lett.*, vol. 81, pp. 30–40, 2016.

[49] J. Yamamoto, K. Inoue, and M. Yoshioka, “Investigation of Customer Behavior Analysis Based on Top-View Depth Camera,” 2017, pp. 67–74.

[50] C. J. Wu, S. Houben, and N. Marquardt, “EagleSense: Tracking people and devices in interactive spaces using real-time top-view depth-sensing,” in *Conference on Human Factors in Computing Systems - Proceedings*, 2017, vol. 2017-May, pp. 3929–3942.

[51] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, “Real Time Head Pose Estimation from Consumer Depth Cameras,” in *Pattern Recognition*, 2011, pp. 101–110.

[52] D. Liciotti, E. Frontoni, A. Mancini, and P. Zingaretti, “Pervasive system for consumer behaviour analysis in retail environments,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10165 LNCS, pp. 12–23.

[53] D. Brazey and C. Gout, “An algorithm for automatic people detection from depth map sequences,” in *EUVIP 2014 - 5th European Workshop on Visual Information*

Processing, 2015.

[54] X. Zhang, J. Yan, S. Feng, Z. Lei, D. Yi, and S. Z. Li, “Water Filling: Unsupervised People Counting via Vertical Kinect Sensor,” 2012.

[55] M. H. Khan, K. Shirahama, M. S. Farid, and M. Grzegorzek, “Multiple human detection in depth images,” in *2016 IEEE 18th International Workshop on Multimedia Signal Processing, MMSP 2016*, 2017.

[56] T. Z. Qiao and S. L. Dai, “Fast head pose estimation using depth data,” in *Proceedings of the 2013 6th International Congress on Image and Signal Processing, CISP 2013*, 2013, vol. 2, pp. 664–669.

[57] D. J. Tan, F. Tombari, and N. Navab, “Real-Time Accurate 3D Head Tracking and Pose Estimation with Consumer RGB-D Cameras,” *Int. J. Comput. Vis.*, vol. 126, no. 2, pp. 158–183, 2018.

[58] D. J. Tan and S. Ilic, “Multi-forest tracker: A Chameleon in tracking,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1202–1209.

[59] D. J. Tan, F. Tombari, S. Ilic, and N. Navab, “A versatile learning-based 3d temporal tracker: Scalable, robust, online,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, vol. 2015 Inter, pp. 693–701.

[60] D. Liciotti, M. Paolanti, E. Frontoni, and P. Zingaretti, “People Detection and Tracking from an RGB-D Camera in Top-View Configuration: Review of Challenges and Applications,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10590 LNCS, pp. 207–218.

[61] L. Zhang, J. Sturm, D. Cremers, and D. Lee, “Real-time human motion tracking using multiple depth cameras,” in *IEEE International Conference on Intelligent Robots and Systems*, 2012, pp. 2389–2395.

[62] L. Chen, H. Wei, and J. Ferryman, “A survey of human motion analysis using depth imagery,” *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 1995–2006, 2013.

[63] E. Murphy-Chutorian and M. M. Trivedi, “Head pose estimation in computer vision: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[64] M. Wolfram, H. Ali, and A. Albu-Schaffer, “Visual Focus of Attention Recognition from Fixed Chair Sitting Postures Using RGB-D Data,” 2017, pp. 325–328.

[65] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, “Random Forests for Real Time 3D Face Analysis,” *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, Feb. 2013.

[66] K. Buys, C. Cagniart, A. Baksheev, T. De Laet, J. De Schutter, and C. Pantofaru, “An adaptable system for RGB-D based human body detection and pose estimation,” *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 39–52, 2014.

[67] A. Brutti and O. Lanz, “A joint particle filter to track the position and head orientation of people using audio visual cues,” in *European Signal Processing Conference*, 2010, pp. 974–978.

[68] M. C. D. F. Macedo, A. L. Apolinário, and A. C. D. S. Souza, “A robust real-time face tracking using head pose estimation for a markerless AR system,” in *Proceedings - 2013 15th Symposium on Virtual and Augmented Reality, SVR 2013*, 2013, pp. 224–227.

[69] Intel, “Intel RealSense Depth Camera D435,” 2019. [Online]. Available: https://ark.intel.com/content/www/us/en/ark/products/128255/intel-realsense-depth-camera-d435.html. [Accessed: 06-Jul-2020].

[70] P. Viola and M. J. Jones, “Robust Real-time Object Detection,” 2001.

[71] “ OpenCV 4.0.” [Online]. Available: https://opencv.org/opencv-4-0/. [Accessed: 24-

May-2019].

[72] “MRC Viewable Ad Impression Measurement Guidelines Prepared in collaboration with IAB Emerging Innovations Task Force Version 1.0 (Final),” 2014.

[73] “Histogram Comparison — OpenCV 2.4.13.7 documentation.” [Online]. Available: https://docs.opencv.org/2.4/doc/tutorials/imgproc/histograms/histogram_comparison/histogram_comparison.html#results. [Accessed: 26-Aug-2020].

[74] K. Saho, “Kalman Filter for Moving Object Tracking: Performance Analysis and Filter Design,” in *Kalman Filters - Theory for Advanced Applications*, InTech, 2018.

[75] B. Feng, M. Fu, H. Ma, Y. Xia, and B. Wang, “Kalman filter with recursive covariance estimation-Sequentially estimating process noise covariance,” *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6253–6263, 2014.

[76] Y. Ioannou, B. Taati, R. Harrap, and M. Greenspan, “Difference of normals as a multi-scale operator in unorganized point clouds,” in *Proceedings - 2nd Joint 3DIM/3DPVT Conference: 3D Imaging, Modeling, Processing, Visualization and Transmission, 3DIMPVT 2012*, 2012, pp. 501–508.

[77] Radu Bogdan Rusu and Steve Cousins, “Point Cloud Library (PCL): PCL API Documentation.” [Online]. Available: http://pointclouds.org/documentation/. [Accessed: 26-Aug-2020].

[78] “Coordenadas homogéneas.” .

[79] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[80] R. Hu, M. Barnard, and J. Collomosse, “Gradient field descriptor for sketch based retrieval and localization,” in *Proceedings - International Conference on Image Processing, ICIP*, 2010, pp. 1025–1028.

[81] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J.*

Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.

[82] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[83] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," 2001.

[84] Y. Freund, Y. Freund, and R. E. Schapire, "A Short Introduction to Boosting," *Proc. Sixt. Int. Jt. Conf. Artif. Intell.*, vol. 14, no. 14, pp. 1401--1406, 1999.

[85] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li, "Learning Multi-scale Block Local Binary Patterns for Face Recognition," in *Advances in Biometrics*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 828–837.

[86] Open Source Robotics Foundation, "ROS/Introduction - ROS Wiki," *ROS Wiki*, 2018. [Online]. Available: http://wiki.ros.org/ROS/Introduction. [Accessed: 06-Aug-2020].

[87] M. Quigley *et al.*, "ROS: an open-source Robot Operating System," *ICRA Work. open source Syst.*, 2009.

[88] Intel Corporation ®, "Intel ® RealSense TM D400 Series Product Family," no. July, pp. 1–114, 2018.

[89] G. F. Torres del Castillo, "La representación de rotaciones mediante cuaterniones," in *Miscelánea Matemática*, no. 29, 1999, pp. 43–50.

[90] "Wikipedia. Cuaterniones y rotación en el espacio." [Online]. Available: https://es.wikipedia.org/wiki/Cuaterniones_y_rotación_en_el_espacio. [Accessed: 06-Aug-2020].

[91] S. Jose, "ICM-20690 Datasheet," vol. 1, no. 408, pp. 1–37, 2015.

[92] I. Caminal, "DETECCIÓ DE PROMINÈNCIA UTILITZANT EL MOVIMENT DE PERSONES AMB VISIÓ ZENITAL," ESEIAAT (UPC), Terrassa, 2018.

[93] M. Lopez-Palma, J. Gago, M. Corbalan, and J. R. Morros, “Audience measurement using a top-view camera and oriented trajectories,” in IECON Proceedings (Industrial Electronics Conference), 2019, vol. 2019-Octob, pp. 74–79.