Universidad de Alcalá
Escuela Politécnica Superior

Grado en Ingeniería Electrónica y Automática Industrial

Trabajo Fin de Grado

Diseño, generación y anotación de base de datos de secuencias de imágenes en interiores para aplicaciones de video-vigilancia

Autor: Valeria Boggian Arévalo

Tutor/es: Cristina Losada Gutiérrez y Marta Marrón Romera

2016

UNIVERSIDAD DE ALCALÁ

ESCUELA POLITÉCNICA SUPERIOR

GRADO EN INGENIERÍA ELECTRÓNICA Y AUTOMÁTICA INDUSTRIAL

TRABAJO FIN DE GRADO

Diseño, generación y anotación de base de datos de secuencias de imágenes en interiores para aplicaciones de video-vigilancia

AUTORA: VALERIA BOGGIAN ARÉVALO

DIRECTOR/ES: CRISTINA LOSADA GUTIÉRREZ

MARTA MARRÓN ROMERA

TRIBUNAL:

PRESIDENTE: MANUEL MAZO QUINTAS

VOCAL 1º: JUAN MANUEL MIGUEL JIMÉNEZ

VOCAL 2º: CRISTINA LOSADA GUTIÉRREZ

FECHA: 16 DE SEPTIEMBRE DE 2016.

CALIFICACIÓN:

Resumen

En el siglo XXI las aplicaciones que engloban el análisis e interpretación de vídeos van adquiriendo cada vez más relevancia debido a sus múltiples aplicaciones.

Para la validación de estas tecnologías se utilizan bases de datos ya publicadas en el ámbito de la investigación, pero estas bases de datos no siempre están orientadas a la aplicación de este proyecto, la video-vigilancia, ni estudian acciones realistas. Por ello, este proyecto tiene como objetivo la creación de una base de datos que permita el reconocimiento de personas y acciones realistas en interiores, dirigido a este tipo de aplicaciones.

Palabras clave: Video-vigilancia, base de datos, reconocimiento humano, reconocimiento de acciones

Abstract

On the 21st century applications which include analysis and interpretation of videos have acquired much importance due to its multiple applications, such as security, analysis of consumer behavior or the assistance to people with disabilities.

In order to study this kind of technology, there are published databases in the field of research that assist in the investigation, so the aim of this project it is to create a new database that allows recognition of people and indoor realistic actions, directed, especially, to video surveillance.

Keywords: Video surveillance, action dataset, human recognition, action recognition

Resumen extendido

Las aplicaciones que engloban el análisis e interpretación de vídeos forman parte del ámbito de lo que se conoce como inteligencia artificial, la cual se centra en el desarrollo de sistemas de procesamiento de datos capaces de imitar a la inteligencia humana, y como la inteligencia del ser humano, se requiere un aprendizaje y la solución de los problemas que se puedan originar.

Por ejemplo, a un niño cuando es pequeño se le enseña en su casa una foto de un elefante y se le dice qué animal es. Este proceso se repite varias veces hasta que el niño sea capaz de identificar él solo cualquier elefante que haya en el mundo, ya sean africanos o asiáticos o bien los vea en una película.

El objetivo de la inteligencia artificial es entrenar a las máquinas para que estas puedan realizar actividades por sí solas, al igual que la mente humana.

Por otro lado, para que las máquinas sean capaces de ser autónomas es usual diseñar algoritmos. Dichos algoritmos deben ser testados para comprobar su correcto funcionamiento y realizar las mejoras oportunas.

Dentro de la inteligencia artificial, el análisis visual de movimiento humano es uno de los terrenos más estudiados e investigados debido a su gran variedad de aplicaciones, como por ejemplo video-vigilancia, aplicaciones de realidad virtual, seguridad, etc.

Habitualmente, los sistemas de reconocimiento de la actividad humana se basan en tres actividades como son el posicionamiento de la persona dentro del escenario estudiado, el seguimiento de la persona en dicho escenario y el reconocimiento de la actividad realizada, ya sea andar, saltar, etc. ([1])

Debido a estas tres actividades que se han estudiado, es habitual entrenar a la máquina que realiza el reconocimiento a través de bases de datos donde se puede realizar una comparación de los resultados obtenidos por dicha máquina (Figura 2) con los datos guardados en la base de datos (Figura 1).

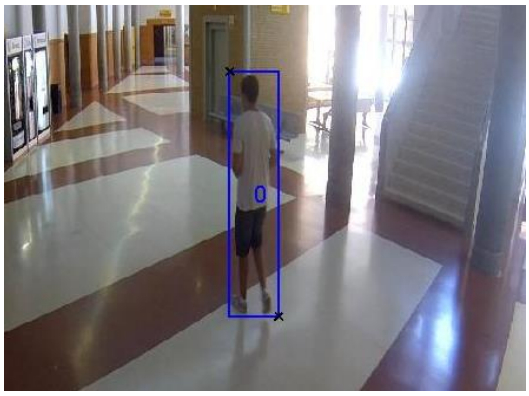


Figura 1 Datos guardados en la base de datos (ideal)

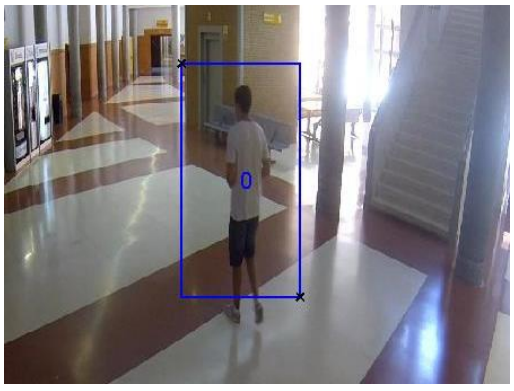


Figura 2 Datos obtenidos por la máquina (real)

Los datos obtenidos en la base de datos, denominados *Ground Truth,* representan el valor ideal que debería generar el sistema y se obtienen de manera manual a través de un proceso denominado anotación.

Lo habitual es que este proceso se realice a través de una interfaz de usuario que permita ir *frame* a *frame* anotando, es decir, indicando dónde está cada persona para así tener mayor fiabilidad en la localización de los puntos.

Como se puede comprobar, el uso de las bases de datos es fundamental tanto para el correcto funcionamiento de cualquier sistema de aprendizaje de una máquina en el campo de la inteligencia artificial, como para la verificación de la funcionalidad de dicha máquina.

Por este motivo en este proyecto se ha creado una base de datos etiquetada que permite la identificación de los usuarios, su seguimiento y la detección de las acciones realizadas. De esta manera se proporciona un buen entrenamiento y buena evaluación de la máquina. Teniendo en cuenta la aplicación de la base de datos, los vídeos que la componen incluyen distintas personas realizando acciones realistas, cotidianas y típicas en aplicaciones de video-vigilancia de interiores tales como andar, sentarse, etc

Además, se realiza una interfaz de anotación capaz de semi-automatizar el proceso de generación de los *Ground Truth*. En estos GT se almacenarán los identificadores de los distintos actores, sus coordenadas y la acción que llevan a cabo, que puede ser andar, correr, sentarse o caerse.

Para crear la interfaz de anotación se ha implementado un código específico para la base de datos desarrollada. Este código permite al usuario realizar distintas acciones dentro del etiquetado, como, por ejemplo, seleccionar el directorio donde se desea trabajar, elegir el identificador del usuario, definir la acción que realiza, establecer el intervalo de *frames* a etiquetar, navegar a lo largo de la anotación a través de los *frames* ya anotados o ver los resultados del proceso.

Debido a que se desea que la interfaz sea intuitiva, se han añadido distintas opciones para que el usuario de la interfaz sea capaz de realizar la anotación de manera sencilla, por ejemplo, a través de cuadros informativos o a través de botones específicos.

Por último, para que los resultados se vean de manera clara, en el etiquetado de los vídeos se puede ver a cada usuario marcado por una caja de anotación de diferentes colores, dependiendo de la acción, y con su número identificador correspondiente.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Capítulo 1

Introducción

Este primer capítulo se emplea para realizar una introducción al proyecto que se presenta además de para explicar los objetivos buscados a la hora de realizar el proyecto, así como la organización y la estructura de la memoria.

1.1. Motivación

El motivo que lleva a realizar y presentar este proyecto es la necesidad de crear una base de datos completa incluyendo acciones humanas que sean realistas, del *día a día*, como por ejemplo sentarse y caerse, cuyos usuarios sean personas de distintas características (mujeres, hombres, mayores, jóvenes, etc.) de tal manera que permita realizar el entrenamiento y evaluación de algoritmos para el reconocimiento de actividades humana.

1.2. Objetivos

El objetivo principal de este proyecto es el diseño, grabación y anotación de una base de datos de imágenes de color, para aplicaciones de video-vigilancia en interiores, además de una interfaz gráfica sobre MATLAB que facilite la anotación de la misma.

Este trabajo, por lo tanto, se compone de la recopilación de características sobre la base de datos a diseñar y la anotación de información de posición y comportamiento.

Los objetivos específicos de este proyecto son:

- Buscar información de bases de datos ampliamente utilizadas, y evaluar la necesidad de crear una nueva base de datos para la aplicación de interés.

- Diseñar la base de datos especificando: número y tamaño de los vídeos, "*set up"* de la grabación, tipología de las escenas, número de usuarios, iteraciones y actividades diferentes a grabar, etc.

- Diseñar y evaluar una interfaz gráfica para realizar la generación de secuencias de vídeo y anotación de personas basándose en una interfaz disponible ([2]), la "*Idiap Head Pose Database"*, con las siguientes características:
    - Interfaz amigable y de fácil uso.
    - Permitirá el etiquetado en intervalos prefijados por el usuario a tantas personas como aparezcan en la secuencia, designándoles a cada una de ellas un identificador diferente y una acción específica.
    - Será capaz de completar la anotación para conseguir que todos los *frames* del vídeo estén cumplimentados correctamente.

  - Aquellos parámetros que puedan ser configurables deberán entenderse de manera sencilla.
  - Se realizará un manual de usuario para facilitar la utilización de la interfaz.

- Integrar en la interfaz la posibilidad de anotar las acciones realizadas por cada usuario.

- Anotar a las personas en los vídeos evaluando, a través de dicha anotación, la base de datos y mejorándola en el ámbito de lo posible.

- Realizar la documentación necesaria tanto para el usuario (manual de usuario) como para el programador.

1.3. Estructura de la memoria

La memoria de este Trabajo fin de Grado está formada por 7 capítulos. A continuación, se indican dichos capítulos y se procede a hacer un breve resumen de su contenido

Capítulo 1. Introducción. (Capítulo actual). En este capítulo se explica los motivos por los cuáles surge este trabajo, así como los objetivos perseguidos y la estructura de la memoria.

Capítulo 2. Marco teórico. Se presenta un estudio sobre las bases de datos utilizadas en el ámbito de la video-vigilancia para entender mejor los motivos por los cuáles se desarrolla la nueva base de datos.

Capítulo 3. Algoritmo de anotación. Capítulo donde se describe el algoritmo de anotación explicando cómo funciona el código de anotación, las distintas funciones del código y de cómo se van desarrollando los archivos GT que permiten realizar el etiquetado.

Capítulo 4. Evaluación. Se presenta la base de datos finalizada junto con la descripción de los vídeos realizados para implementarla.

Capítulo 5. Conclusiones y trabajos futuros. Se exponen las conclusiones obtenidas al finalizar el trabajo y los posibles trabajos futuros.

Apéndice A. Presupuesto. En este capítulo se calcula y presenta el presupuesto asociado al trabajo realizado.

Apéndice B. Manual de usuario. En este apéndice se presenta el manual de usuario de la interfaz gráfica de anotación.

Capítulo 2

Marco teórico

Desde los inicios en la investigación sobre la visión artificial, y por tanto, en la investigación sobre el reconocimiento de la actividad humana, ha sido necesario realizar evaluaciones sobre los algoritmos implementados para comprobar su correcto funcionamiento, así como para comparar los distintos trabajos realizados. ([3])

Además, en las muchas alternativas existentes, es necesario realizar una fase de entrenamiento del sistema para el cual también es necesario la realización de una base de datos completa que abarque todas las posibilidades para que, en un futuro, la máquina esté adiestrada para poder actuar en cualquier situación.

A lo largo del avance en este ámbito de la investigación han sido publicadas una gran cantidad de bases de datos con el objetivo de realizar una identificación de personas y acciones de todo tipo, tanto en interiores como en exteriores, acciones cotidianas o acciones más específicas, usuarios masculinos, femeninos o mixtos, interacciones entre grupos y personas, etc.

En este capítulo se va a realizar un breve estudio sobre la metodología de grabación, las acciones estudiadas, las bases de datos más importantes y/o utilizadas, y las escenas de grabación.

2.1. Metodología de grabación

La metodología de grabación se refiere al método de captura de los datos: vídeo, mapas de profundidad, seguimiento de la figura esquelética, captura de movimiento a través del seguimiento de marcadores (MoCap), unidades de medición inercial (IMU), etc. ([4]).

2.1.1. Vídeo

Las colecciones de imágenes para realizar las bases de datos son capturadas a través de cámaras, normalmente comerciales. Este método es bastante utilizado ya que una cámara proporciona un método de grabación relativamente económico y fácil de utilizar, además de ser muy adaptable a todo tipo de situaciones debido a que permite la grabación en interior, en exterior, con diferentes puntos de vista, etc.

Es posible encontrar distintas opciones cuando se aplica esta metodología de grabación, como por ejemplo la utilización de cámaras estáticas, cámaras móviles o un sistema de múltiples cámaras. El uso de cámaras estáticas es el método más sencillo para obtener la información deseada, aunque bien es cierto que es posible que se pierdan datos,

por ejemplo, por oclusión. Por otro lado, una cámara móvil permite centrar el seguimiento en un objeto específico, pero esto requiere algoritmos más complejos y costosos, y un sistema de múltiples cámaras, que si bien podría resolver los problemas de oclusión, es un método demasiado complejo en cuanto a la instalación de los equipos y a la coordinación de las cámaras. ([1])

En algunos casos, la calidad de las grabaciones puede no ser del todo óptima debido al tipo de cámara, al movimiento de esta, a la pérdida de secuenciase e información debido a la situación de la cámara en la escena, etc.

Por todos estos motivos es necesario realizar un análisis, en el caso de que se elija este método de grabación, para realizar las grabaciones de la mejor manera posible.

2.1.2. Motion Capture

La captura de movimiento, *Motion Capture (MoCap)*, como se puede ver en la Figura 3, es una técnica de grabación que permite capturar los movimientos en el mundo real, a través de distintos marcadores colocados por todo el cuerpo, y trasladar los datos a un modelo virtual ([5]). Este método incluye sistemas mecánicos, acústicos, ópticos e incluso magnéticos.

Figura 3 Captura de movimiento para la película Avatar ([22])

Estos sistemas son muy fiables y permiten capturar el movimiento completo incluyendo las tres dimensiones, siendo así más fácil su identificación, sin embargo, el utilizar los marcadores en los cuerpos de los actores hace que los movimientos no sean tan fluidos y naturales, además de tratarse de un método altamente invasivo.

2.1.3. Mapas de profundidad

Esta técnica permite capturar los movimientos a través de un sensor de profundidad, normalmente la Microsoft Kinect.

Este método permite obtener las posiciones relativas de los puntos más representativos de la imagen real, obtenidos a partir de infrarrojos, para calcular la profundidad de estos y obtener la posición de cada pixel en la imagen virtual creada ([6]), obteniendo así información en 3D.

Figura 4 Imágenes reales y las imágenes obtenidas a través de Kinect ([6])

2.2. Acciones estudiadas

Para la creación de las bases de datos es necesario elegir previamente las acciones que van a ser estudiadas dependiendo del objetivo de la base de datos que se va a crear.

Dentro del ámbito de la investigación se utiliza una terminología determinada para referirse a las distintas clases de acciones, dependiendo de su complejidad o del número de personas que interactúen:

- **Pose:** una pose describe la disposición espacial de un cuerpo humano en un solo momento temporal. Es una postura momentánea y poco natural. Solo implica a una persona.

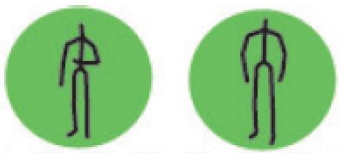


Figura 5 Pose mano subida (izquierda), pone natural (derecha) ([4])

- **Gesto:** Movimiento, normalmente de una parte del cuerpo, formado por una serie de poses de breve duración. Son las acciones primitivas.

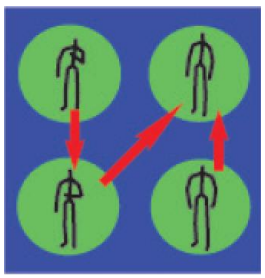


Figura 6 Gesto de subir y bajar la mano ([4])

- **Acción:** Conjunto de gestos que implican movimiento y el cambio de estado de una persona, como, por ejemplo, andar, correr, saltar, etc. En el análisis de la actividad humana son las clases mayor estudiadas debido a su moderada complejidad ya que son aquellas que más realizamos.

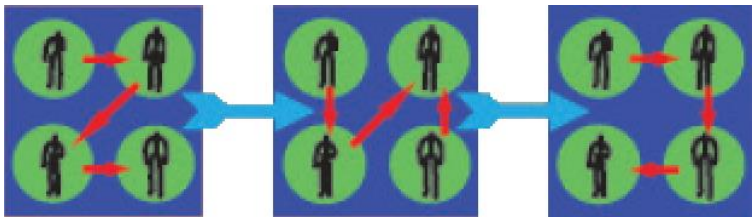


Figura 7 Acción conformada por gestos ([4])

- **Interacción:** Acción recíproca entre dos personas. Cada una de esas personas realiza una acción individual, que, comparada con la acción de la otra persona, surge la interacción. Esta clase de acciones se puede dar entre personas, como por ejemplo el saludarse, o entre persona-objeto, como por ejemplo coger una silla. También se puede realizar el estudio de interacciones entre grupos de más de dos personas

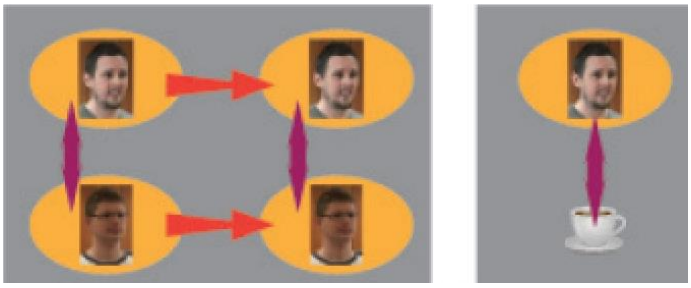


Figura 8 Interacción entre personas (izq.) o persona-objeto (der.) ([4])

- **Actividad:** Una actividad es un conjunto de acciones/interacciones de alta complejidad que conforman un evento de alto nivel. Las actividades más estudiadas son las cotidianas, como por ejemplo hacer la cama o interacciones como coger una sábana.

Figura 9 Actividad que desarrollan al hacer café ([4])

2.3. Ground Truth

El término *Ground Truth* (GT) se refiere al proceso de recopilación de datos específicos, es decir, cuando se crea una base de datos y se trabaja con ella se generan unos archivos resultantes de la anotación de la base de datos donde se recoge la información que desea ser estudiada, como por ejemplo la posición de la persona y la acción desarrollada por la misma.

Cuando se crean bases de datos para este tipo de aplicaciones estas están compuestas por una serie de vídeos formados por los primeros planos, el fondo y elementos para reconocer. El proceso de anotación o etiquetado determina qué elementos presentes en las imágenes se desean estudiar y qué información se desea obtener sobre estos, por lo que las bases de datos deben contener las imágenes etiquetadas con sus archivos GT asociados.

En algunas bases de datos se realizan anotaciones de los GT de manera manual *frame a frame* mientras que otras se realizan el anotado de manera semi-automatizada, es decir, cada cierto número de *frames* o de secuencia en secuencia de manera manual siendo completado de manera automática a través de un algoritmo específico.

Cuando se realiza el proceso de anotación se deben tener en cuenta:

- La región de interés en la escena de grabación, es decir, la zona se desea realizar el estudio de reconocimiento.
- El contenido de la escena, teniendo en cuenta los elementos fijos como por ejemplo las columnas; o las variables dinámicas, como las personas que puedan aparecer.
- La luz de la escena de grabación, ya que en ciertos casos puede verse afectado el proceso de anotación debido a cambios importantes en iluminación
- Las referencias o los ejes de referencia de nuestro sistema
- Las posibles oclusiones de las personas de interés

Cada base de datos tiene su GT personal ya que cada una de ellas obtiene un tipo de información de los vídeos. Por ejemplo, en la base de datos ARENA se realiza un seguimiento de objetos, determinando en qué lugar se encuentra en cada momento el objeto a seguir. Así mismo, se detecta qué tipo de objeto es el que se está siguiendo, si un vehículo o una persona, y por último se realiza una anotación de la actividad desarrollada, por ejemplo, andar ([21]).

Otro ejemplo sería la base de datos BEHAVE, donde los archivos GT tienen almacenados el identificador de grupo al que pertenecen, el identificador del otro grupo con el que se está interactuando (ver apartado 2.4.2), el *frame* en el que comienza la acción, el *frame* en el que finaliza dicha acción y, por último, la acción realizada. ([8])

Los *ground truth* pueden proporcionar un claro punto de referencia para aprobar o refutar los métodos utilizados para el reconocimiento de actividad humana.

2.4. Bases de datos

Como ya se ha dicho anteriormente, existen una multitud de bases de datos utilizadas para la investigación en el ámbito del reconocimiento humano, con distintas acciones estudiadas, diferentes escenas de grabación y utilizando distintas metodologías.

Algunas de las bases de datos a las que se suele hacer referencia en los trabajos analizados para hacer este proyecto se muestran en la Tabla 1, resaltando las más utilizadas ([3]):

Tabla 1 Bases de datos más utilizadas en el reconocimiento de actividad humana

Arena Dataset	CROSSING	LINTHESCHER	PlaceLab datasets
Bahnhof	**Human Eva I**	LOEWENPLATZ	SUNNY DAY
BEHAVE Dataset	IDIAP	MHAD	TownCentre Dataset
Berkeley MHAD	i-LIDS	MuHAVI dataset	TUD
Cambridge-Gestures	**INRIA**	MSR II	TUM
CAVIAR	**IXMAS**	PARADEPLATZ	VIRAT Human Activity
Cha Learn	JELMOLI	PEDCROSS 2	**Weizmann**
CMU	**KTH Human Motion**	PETS	…

De las bases de datos mostradas en la Tabla 1 se va a realizar un breve estudio de las marcadas en negrita.

2.4.1. ARENA DATASET

Base de datos con veintidós vídeos creada para el proyecto ARENA (*Arquitecture for the REcognition of threats to mobile assets using Networks of multiple Affordable sensors)* en la Universidad de Reading, Berkshire, Inglaterra. ([7])

El objetivo principal de esta base de datos es la detección y el análisis del comportamiento humano en un parking con un vehículo aparcado.

ESCENA DE GRABACIÓN

Las grabaciones están realizadas en el aparcamiento situado en frente de la Escuela de Sistemas de Ingeniería, de tal manera que se colocan cuatro cámaras ambientales para cubrir un área aproximada de 100mx30m. A través de estas cámaras se obtiene una visión global del área monitorizada, como se puede comprobar en la Figura 10 y en la Figura 11.

Las cámaras ambientales utilizadas son de dos modelos diferentes, pero teniendo ambas las mismas prestaciones:

- Resolución: 768x576 píxeles
- Frame rate: 7fps
- Escáner progresivo

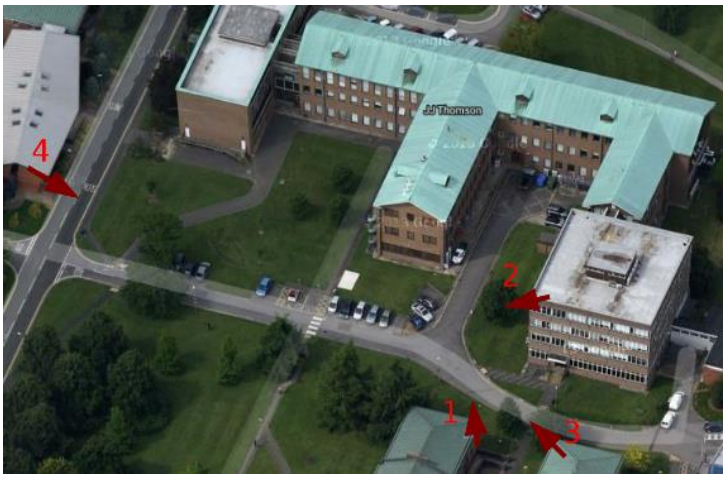


Figura 10 Instalación de las cámaras para la grabación de la base de datos ARENA ([7])

Figura 11 Puntos de vista de las cámaras de la base de datos ARENA ([7])

Además de las cámaras ambientales se tienen cuatro cámaras colocadas en cada esquina de un vehículo aparcado en la zona de visión con las siguientes características:

- Resolución: 1280x960 píxeles
- Frame rate: 30fps

De esta manera el campo de visión de las cámaras, y, por ende, la región de interés es la siguiente:

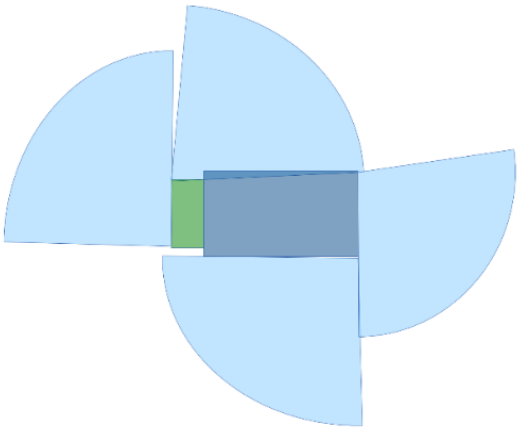


Figura 12 Región de interés de la base de datos ARENA ([7])

ACCIONES ESTUDIADAS

Se realiza el análisis del comportamiento humano en base a los distintos sucesos que puedan ocurrir en este escenario y su grado de "normalidad" teniendo las siguientes categorías:

- **Situaciones anormales**: situaciones importantes de analizar pero que no suponen amenaza
  - Una persona colocada de pie y sin moverse al lado del vehículo
  - Una persona o varias personas andando alrededor del vehículo
  - El conductor del vehículo se cae del mismo y una persona o varias van a ayudarle
  - Una persona pregunta una dirección al conductor del vehículo
- **Situaciones de riesgo**: situaciones en las que la seguridad del conductor, del vehículo o de la carga del vehículo están bajo amenaza
  - Una persona andando alrededor del vehículo intenta abrirlo
  - El conductor se cae del vehículo empujado por alguien
- **Situaciones críticas**: situaciones en las que la seguridad del conductor, del vehículo o de la carga del vehículo están en peligro
  - Una persona o un grupo de personas robando el vehículo
  - El conductor es agredido por una persona o por un grupo de personas
  - El conductor es golpeado y la fuga de los agresores se da en coche

2.4.2. BEHAVE DATASET

Base de datos conformada por cuatro vídeos creada para el Instituto de la Percepción de Acciones y Actividad Humana, en la Universidad de Edimburgo, Escocia, Reino Unido. ([8])

El objetivo principal es la detección y el análisis del comportamiento humano en una calle con vehículos aparcados en una zona no peatonal.

ESCENA DE GRABACIÓN

Los cuatro vídeos son grabados en el mismo escenario, una calle no peatonal con coches en uno de los laterales. Dos de los vídeos están grabados con una cámara con un ángulo de visión perpendicular a la calle y los otros dos vídeos están grabados con un ángulo de visión paralelo a la misma.

Figura 13 Visión de los puntos de vista de grabación de la base de datos BEHAVE ([8])

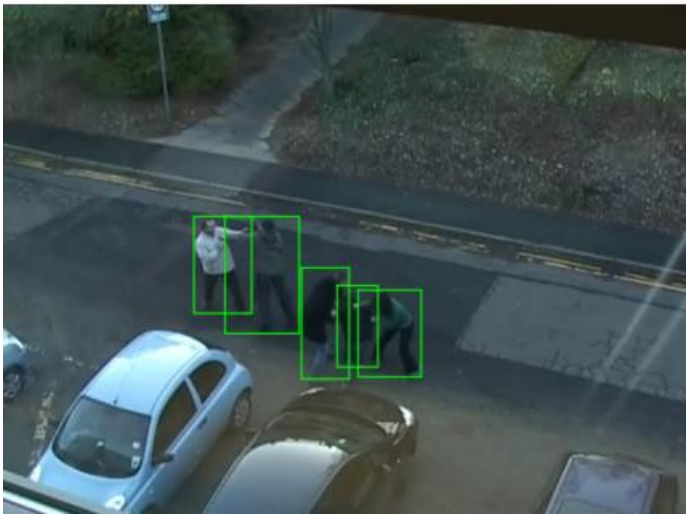


Figura 14 Ejemplo de anotación de un frame de la base de datos BEHAVE ([8])

Las cámaras utilizadas para realizar las grabaciones tienen las siguientes características:

- Resolución: 640x480 píxeles
- Frame rate: 25fps
- Formato de grabación RGB

ACCIONES ESTUDIADAS

La base de datos consiste en un análisis basado en el flujo de comportamiento de varios individuos interactuando, es decir, de dos a cinco personas que interactúan como un grupo o bien como dos grupos que interactúan entre sí.

Esta base de datos es creada para la detección de situaciones delictivas o peligrosas entre interacciones con grupos pequeños y situaciones donde se den multitudes en la calle.

De esta manera se obtienen diez tipos de comportamientos de grupo que fueron anotados por el equipo de investigación, los cuales pueden verse en la Tabla 2.

Tabla 2 Listado de interacciones de la base de datos BEHAVE

TIPO DE COMPORTAMIENTO	DESCRIPCIÓN	CASOS DE EJEMPLO EN LOS VÍDEOS
En grupo	Personas en un grupo sin moverse demasiado	35
Aproximación	Dos personas o grupos aproximándose entre ellos	25
Andar Juntos	Personas andando juntas	43
Encontrarse	Una o dos personas encontrándose a otra	1
Separarse	Una o dos personas separándose de otra	23
Ignorar	Una o dos personas ignorando a otra	2
Persecución	Un grupo persiguiendo a otro	10
Pelear	Dos o más grupos peleándose	19
Correr juntos	Un grupo corriendo a la vez	4
Seguir	Un grupo siguiendo a una persona	1

2.4.3. BERKELEY MHAD

Base de datos compuesta por once acciones desarrolladas por 7 individuos diferentes creada para un proyecto de investigación llamado *Teleimmersion Lab* en la Universidad de Berkeley, California, Estados Unidos. ([9])

El objetivo es el reconocimiento de movimientos en varios humanos con distintas características.

ESCENA DE GRABACIÓN

La base de datos Berkeley MHAD *(Berkeley Multimodal Human Action Database)* se basa en la grabación de una persona realizando distintos movimientos.

La acción se desarrolla en una habitación con un único cambio de fondo, en presencia o ausencia de una silla.

Para realizar la adquisición de datos se realiza el montaje audiovisual de la Figura 15.

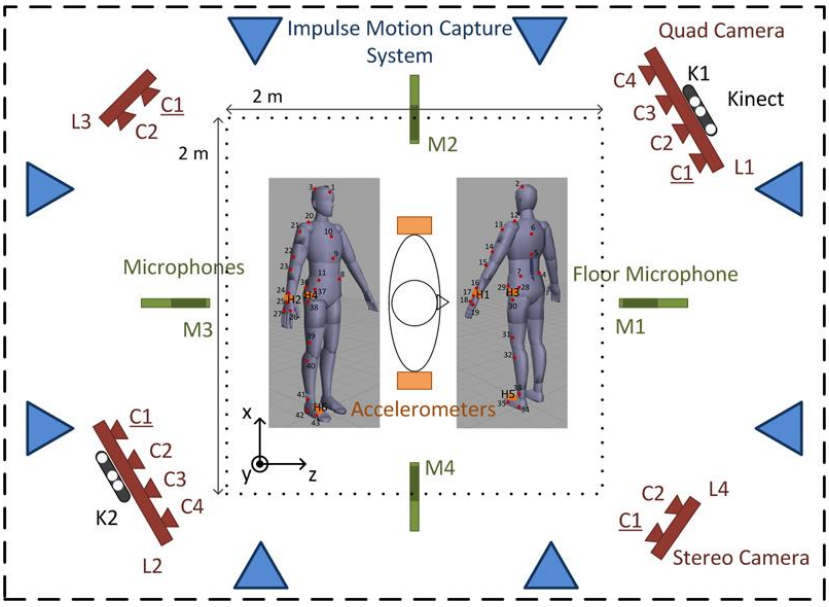


Figura 15 Sistema de grabación de la base de datos BERKELEY MHAD ([9])

Cada acción es capturada simultáneamente por cinco sistemas diferentes:

- Un sistema de captura de movimiento óptico conformado por 12 cámaras (Figura 16) DragonFly con una resolución de 640x480 píxeles y un *frame rate* de 11fps, colocadas en dos grupos, un grupo con dos cámaras para visión estéreo (L3 y L4 en Figura 15) y otros dos grupos formados por cuatro cámaras cada uno (L1 y L2 en Figura 15) para una captura multi-vista.

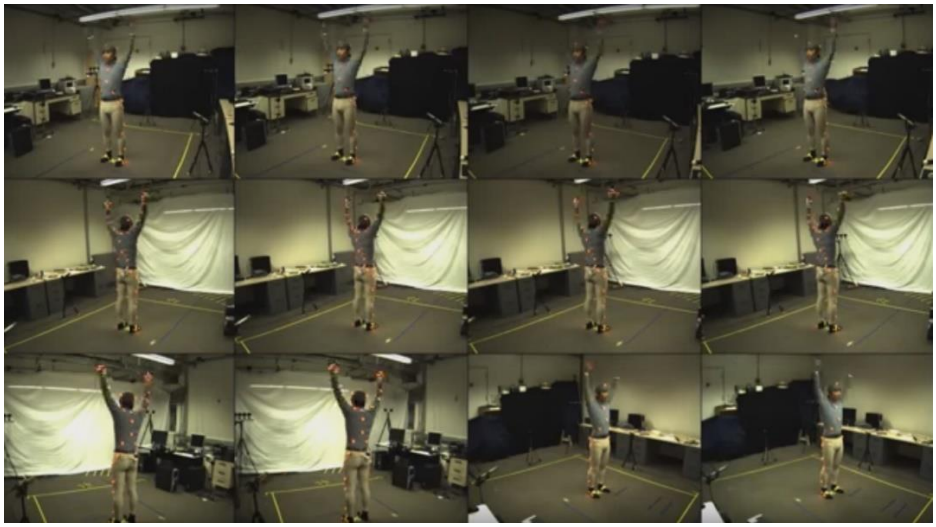


Figura 16 Visión de las 12 cámaras de la base de datos BERKELEY ([9])

- Un sistema de captura de profundidad (ver Figura 17) conformado por dos cámaras de Microsoft Kinect con una resolución de 640x480 pixeles y un *frame rate* de 15fps. Estas están colocadas en posiciones opuestas para prevenir las interferencias entre ellas.

Figura 17 Visión de la Kinect en la base de datos BERKELEY ([9])

- Un sistema formado por seis acelerómetros de tres ejes para medir la dinámica de los movimientos realizados por las muñecas, los tobillos y las caderas del usuario al que se está estudiando.

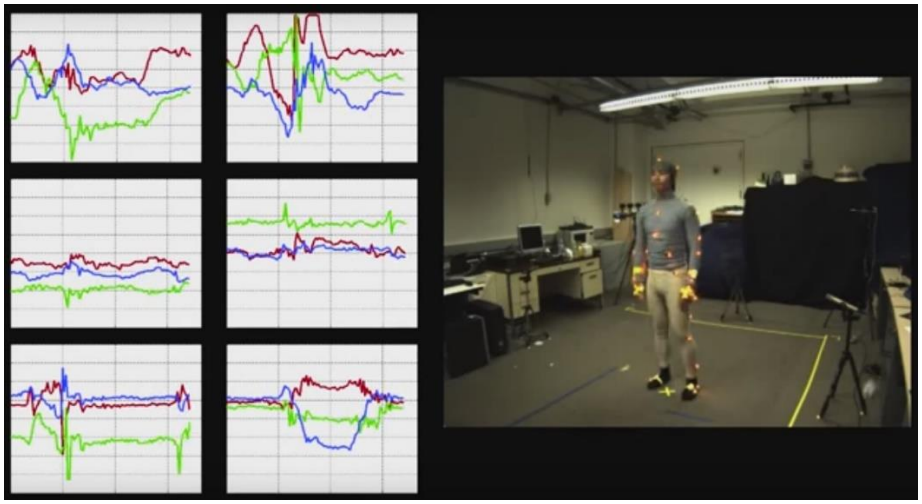


Figura 18 Resultado de los acelerómetros en la base de datos BERKELEY ([9])

- Un sistema formado por unos marcadores LED capaces de capturar movimiento 3D (MoCap) para realizar los archivos GT. Este sistema utiliza unos detectores lineales basados en 8 cámaras con una resolución de 3600x3600 píxeles colocadas alrededor del usuario (ver Figura 15)

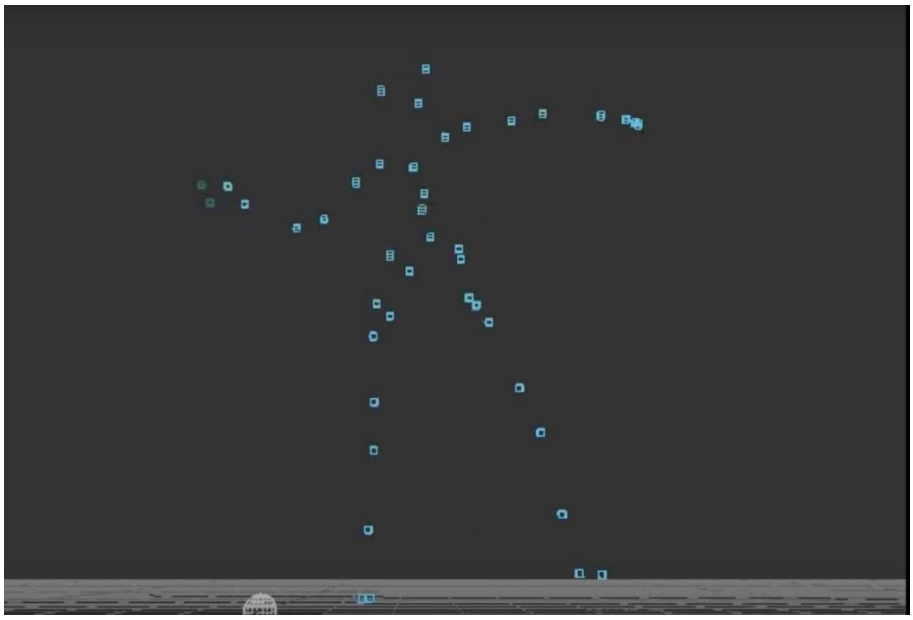


Figura 19 Ejemplo de MoCap de la base de datos BERKELEY ([9])

- Un sistema de audio formado por cuatro micrófonos, tres de los cuales se encuentran sujetos a un trípode para obtener los sonidos del usuario, y uno de los micrófonos colocado en el suelo para registrar los sonidos y vibraciones producidos por este cuando se realizan distintas acciones.

ACCIONES ESTUDIADAS

Las personas elegidas son siete hombres y cinco mujeres de entre 23-30 años realizando cada uno de ellos, y por separado, una acción específica y con unas cinco repeticiones de la misma.

Tabla 3 Acciones estudiadas en la base de datos BERKELEY

ACCIÓN	Nº REPETICIONES	TIEMPO
Saltar en el sitio	5	5 segundos
Salto de tijera	5	7 segundos
Doblarse	5	12 segundos
Boxear	5	10 segundos
Palmada sobre la cabeza	5	7 segundos
Subir la mano derecha sobre la cabeza	5	7 segundos
Aplaudir	5	5 segundos
Lanzar una pelota	1	3 segundos
Sentarse y levantarse	5	15 segundos
Sentarse	1	2 segundos
Levantarse	1	2 segundos

Este conjunto de actividades contiene movimientos corporales dinámicos en general. Algunas actividades tienen una dinámica en ambas extremidades superiores e inferiores mientras que otras actividades tienen dinámicas más en extremidades superiores.

Este conjunto de actividades permite recoger datos de movimiento naturales, ya que los sujetos no fueron instruidos específicamente para realizar cada acción. La espontaneidad es una característica crítica para la captura de diferentes estilos de movimiento para la misma acción en la base de datos.

2.4.4. CAVIAR

Base de datos creada por la Universidad de Edimburgo en colaboración con el Instituto Superior Técnico de Lisboa basada en varios videoclips grabados en distintos escenarios de interés para conseguir distintos objetivos, como la video vigilancia en un centro comercial o el análisis del comportamiento humano en clientes potenciales. ([10])

ESCENA DE GRABACIÓN

Esta base de datos está basada en dos grupos de vídeos, algunos de ellos grabados en el vestíbulo de la entrada a los laboratorios INRIA en Grenoble, Francia, y otro grupo de vídeos grabados en un centro comercial en Lisboa, Portugal.

Figura 20 Vídeos grabados en INRIA para la base de datos CAVIAR ([10])

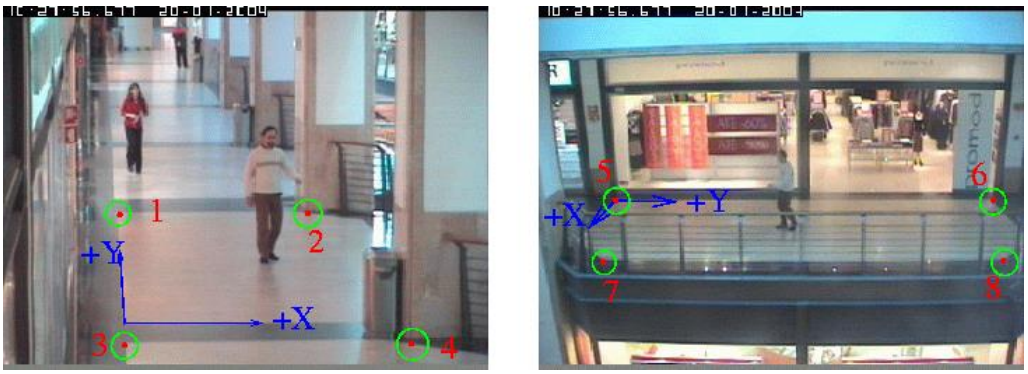


Figura 21 Centro comercial de Portugal para la base de datos CAVIAR ([10])

En el primer set, la grabación está hecha con un ángulo picado desde la esquina superior derecha, y en el segundo set se utilizan dos cámaras, una en posición frontal con respecto a una tienda y la otra en posición longitudinal en el pasillo donde se encuentra dicha tienda. (Ver Figura 20 y Figura 21).

Toda la base de datos está rodada a través de una cámara de lente gran angular con una resolución de 384x288 píxeles a 25fps, con un formato de grabación en RGB.

ACCIONES ESTUDIADAS

Primer Set. Laboratorios INRIA, Francia

La anotación de este primer set se realiza a través de dos identificaciones diferentes, los cuadros amarillos identifican a las personas individualmente, y los cuadros verdes identifican a grupos de personas para estudiar sus interacciones. Los individuos que no son de interés no son reconocidos (anotados) por el sistema.

Figura 22 Anotación del primer set en la base de datos CAVIAR ([10])

Las secuencias obtenidas son las mostradas en la Tabla 4.

Tabla 4 Acciones estudiadas en el primer set de la base de datos CAVIAR

ACCIÓN	Nº VÍDEOS CON LA ACTIVIDAD REALIZADAD
Andar	3
Buscar	6
Sentarse y caerse	4
Dejar una mochila	5
Andar en grupo	6
Dos personas peleándose	4

Segundo Set. Centro Comercial, Lisboa

En este set las grabaciones están realizadas con personas entrando y saliendo de las tiendas, las cuales no son anotadas por la base de datos ya que no son personas de interés.

En este caso las anotaciones se realizan individualmente, identificando, además, la cabeza, los hombros, el tronco las manos y los pies.

Figura 23 Anotación del segundo set en la base de datos CAVIAR ([10])

Las secuencias obtenidas son las mostradas en la Tabla 5.

Tabla 5 Acciones estudiadas en el segundo set de la base de datos CAVIAR ([10])

ACCIÓN	Nº VÍDEOS CON LA ACTIVIDAD REALIZADAS
Pareja andando a lo largo del pasillo	2
Pareja entrando a una tienda a través del pasillo	2
Pareja saliendo de una tienda a través del pasillo	2
Persona saliendo de una tienda	3
Pareja andando por el pasillo, una persona entra en la tienda, la otra espera fuera	2
Pareja entrando y saliendo en una tienda	7
Persona buscando algo o a alguien en una tienda	2
Tres personas andando por el pasillo	2

2.4.5. HUMAN EVA I

Base de datos que estudia el movimiento de cuatro personas realizando seis acciones diferentes. El objetivo es la evaluación del movimiento humano articulado a través de un sistema de captura de movimiento creada por la Universidad de Toronto, Ontario, Canadá en colaboración con la Universidad Brown, Rhode Island, Estados Unidos. ([11])

ESCENA DE GRABACIÓN

La base de datos consta de seis acciones a realizar grabadas tres veces, dos de ellas con vídeo y captura de movimiento y una vez únicamente con captura de movimiento.

La grabación está desarrollada de manera unipersonal en una sala con un fondo fijo.

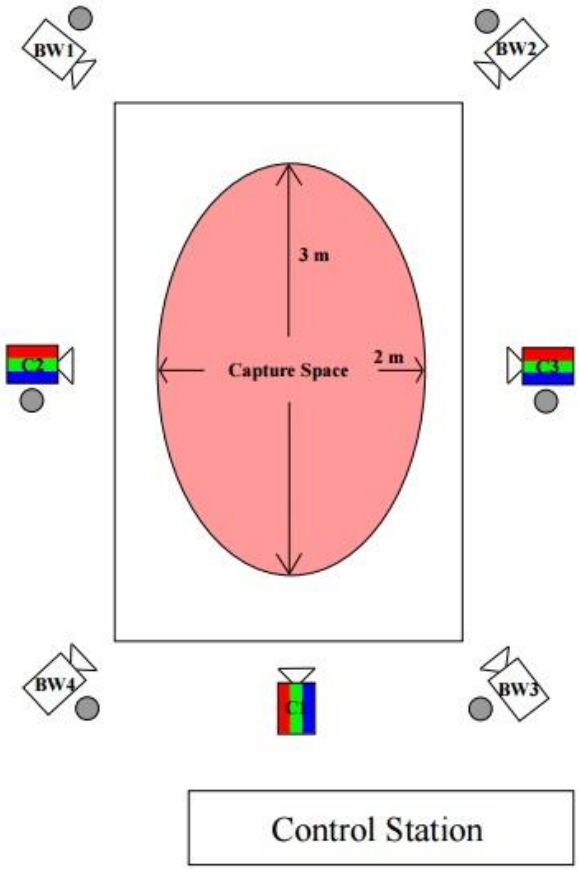


Figura 24 Sistema de grabación de la base de datos Human Eva I ([11])

Para realizar la captura de movimiento se utiliza un sistema basado en marcadores reflectantes de captura de movimiento (MoCap) y seis cámaras para recuperar la posición 3D de los marcadores.

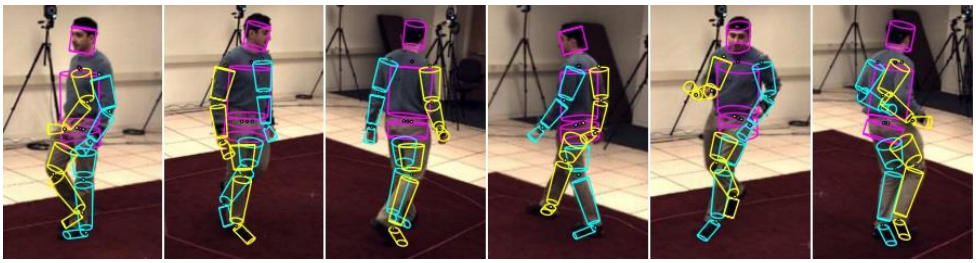


Figura 25 Captura de movimiento de la base de datos HUMAN EVA I ([11])

Las grabaciones de vídeo se realizan utilizando dos sistemas de captura de vídeo comerciales:

- Uno de ellos conformado por cuatro cámaras de escala de grises con una resolución de 648x484 píxeles y un *frame rate* de 120fps (BWX en la Figura 26).

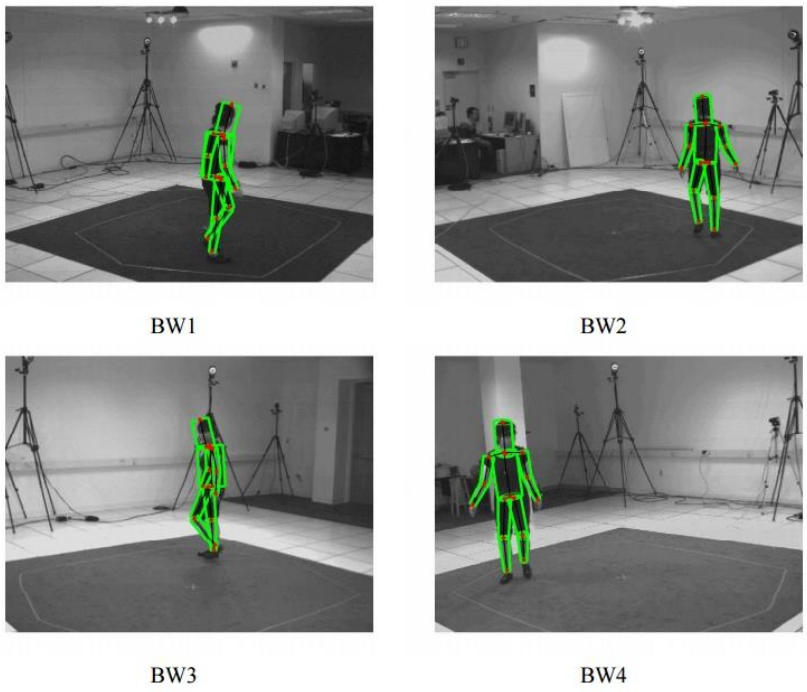


Figura 26 Cámaras en escala de grises de la base de datos HUMAN EVA I ([11])

- El otro sistema está formado por tres cámaras RGB con una resolución de 659x494 píxeles (las cuales fueron modificadas para esta aplicación a 640x480 píxeles) y 55fps (CX en la Figura 27).

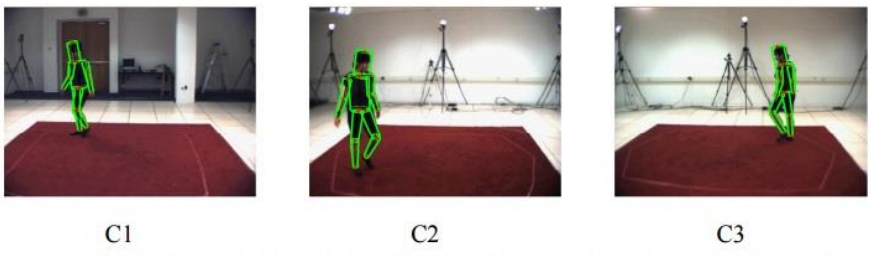


Figura 27 Cámaras RGB para la base de datos HUMAN EVA I ([11])

ACCIONES ESTUDIADAS

Las personas elegidas fueron 4 sujetos, una mujer y tres hombres, realizando, por separado, una acción específica. El conjunto de actividades desarrolladas permite recoger datos de movimiento naturales, ya que los sujetos no fueron instruidos específicamente cómo realizar cada acción.

En cada vídeo se ve a cada una de las personas realizando la acción:

Tabla 6 Acciones estudiadas en la base de datos HUMAN EVA I

ACCIÓN	Nº REPETICIONES
Andar	12
Hacer jogging	12
Realizar el movimiento de hola y adiós con la mano	12
Lanzar y coger una pelota	12
Boxear	12
Combo de acciones: caminar seguido de trotar y luego mantener el equilibrio sobre cada uno de los dos pies	12

2.4.6. IXMAS DATASET

La base de datos IXMAS *(Inria Xmas Motion Acquisition Sequences)* consiste en diez vídeos con distintas acciones con el objetivo de establecer una investigación sobre la captura de información en imágenes con movimiento. ([12][13])

ESCENA DE GRABACIÓN

Para la adquisición de las imágenes se utiliza un escenario de grabación fijo donde cada actor, por separado, realiza las secuencias de acciones prefijadas.

Para la realización de la grabación se utilizan cinco cámaras RGB colocadas en distintos puntos de la sala de grabación con una resolución de 390x291 píxeles y grabadas a 23fps.

Figura 28 Distintos puntos de vista de la base de datos IXMAS ([13])

Las cámaras están colocadas de tal manera que se obtienen 5 perspectivas distintas de la acción:

- Frontal
- Lateral
- Vista superior derecha
- Vista superior izquierda
- Vista superior vertical

ACCIONES ESTUDIADAS

Los diez vídeos que conforman esta base de datos se componen de diez actores (un vídeo con todas las acciones por cada persona), cinco hombres y cinco mujeres, realizando once acciones, tres veces cada acción, de manera individual.

Las acciones realizadas son las siguientes:

- Mirar el reloj de pulsera
- Cruzar los brazos
- Rascarse la cabeza
- Sentarse
- Levantarse
- Girar
- Andar
- Sacudir las manos
- Pegar un puñetazo
- Dar una patada
- Coger algo

Las acciones son grabadas de manera espontánea y no se desarrollan siempre con la persona en el mismo sitio ni con la misma orientación.

2.4.7. KTH HUMAN MOTION DATASET

Base de datos conformada por seiscientos vídeos estudiando seis tipos de acciones creada por el Real Instituto de Tecnología, Estocolmo, Suecia

Es considerada la base de secuencias de acciones pública de mayor tamaño por lo que es la base de datos estándar en el análisis del comportamiento humano. ([14] [15])

ESCENA DE GRABACIÓN

La base de datos está desarrollada en cuatro escenarios diferentes:

- Al aire libre
- Al aire libre con variación de escala
- Al aire libre con cambio de vestuario del actor
- En el interior de una habitación con el fondo blanco.

Todas las secuencias (en total 2391 secuencias), en blanco y negro, son tomadas sobre fondos homogéneos a través de una cámara estática con una resolución de 160x120 píxeles, a una velocidad de 25fps y con una técnica de grabación en RGB.

Figura 29 Ejemplos de secuencias de la base de datos KTH ([14])

Esta base de datos contiene 192 vídeos para la fase de entrenamiento de la máquina de reconocimiento humano, 192 para la fase de validación del algoritmo y 216 vídeos para la fase de pruebas.

ACCIONES ESTUDIADAS

Las acciones desarrolladas en las secuencias son seis acciones humanas unipersonales rodadas por veinticinco sujetos, ocho personas para la fase de entrenamiento, ocho personas para la fase de validación y nueve personas para la fase de test.

Las acciones desarrolladas son:

- Andar
- Hacer jogging
- Correr
- Boxear
- Sacudir las manos
- Aplaudir

2.4.8. WEIZMANN DATASET

La base de datos Wiezmann fue creada por la Universidad Machon Weizmann, Rehovot, Israel, para la verificación de un nuevo algoritmo de detección de actividad humana. ([16])

ESCENA DE GRABACIÓN

Esta base de datos se desarrolla en dos experimentos diferenciados.

La primera fase, o los primeros experimentos, son grabados en baja resolución a través de una cámara estática de 180x144 píxeles a 25fps colocada frente al actor que realiza las acciones.

El escenario de grabación es un fondo homogéneo, en exteriores, realizando la acción de manera individual. En este tipo de grabaciones no hay distintos enfoques ni tampoco posibilidad de interacción o aparición de nuevas personas en la imagen.

Figura 30 Ejemplos de los primeros vídeos de la base de datos WEIZMANN ([16])

La segunda fase de experimentación conlleva una serie de vídeos, diez en total, grabados con distintos fondos diferentes no homogéneos, con las mismas características de grabación. En este caso la diferencia con la primera fase son las acciones estudiadas y las variaciones en los puntos de vista (0^o, 5^o, 10^o, …, 45^o)

Figura 31 Primer fondo de la segunda fase en la base de datos WEIZMANN ([16])

Figura 32 Segundo fondo de la segunda fase en la base de datos WEIZMANN ([16])

Figura 33 Tercer fondo de la segunda fase en la base de datos WEIZMANN ([16])

ACCIONES ESTUDIADAS

En la primera fase las nueve acciones, representadas todas ellas por nueve actores, tanto hombres como mujeres, son:

- Andar
- Correr
- Ir saltando
- Galopar
- Saludar con una mano
- Saludar con las dos manos
- Salto en el sitio
- Saltar a la comba
- Agacharse

Todas estas acciones se pueden considerar acciones diarias y naturales. Para la segunda fase se realiza un estudio sobre una sola acción de distintas maneras, es decir, la acción de andar combinándola con diferentes elementos:

- Andar normal
- Andar con falda
- Andar llevando un maletín
- Andar con cojera
- Andar con las piernas ocultas
- Andar subiendo las rodillas al pecho
- Andar con un perro
- Andar como un sonámbulo
- Andar agitando una bolsa
- Andar en diagonal

2.5. Conclusiones del marco teórico

En primer lugar, tras el breve estudio realizado para dar comienzo a este trabajo, se observa la cantidad de proyectos de investigación, estudios y publicaciones en torno al reconocimiento de la actividad humana.

Se comprueba que es una parte de la inteligencia artificial que en estos momentos se encuentra en auge y la cual es de gran interés.

Tras lo expuesto en el marco teórico, se puede realizar una comparación de las bases de datos analizadas.

Tabla 7 Comparación entre las distintas bases de datos

Base de datos	Método de grabación	Clase de acción	Tamaño			Aplicación	Escena de grabación	
			Acciones	Personas	Secuencias		Mono/multivista	Entorno
ARENA Dataset	RGB	Acción Interacción	11 - 15	Indefinido	Indefinido	Video-vigilancia	Multivista	Exterior
BEHAVE Dataset	RGB	Acción Interacción	6 - 10	Indefinido	< 20	Video-vigilancia	Multivista	Exterior
BERKELEY MHAD	RGB-D, IMU, Audio, MoCap	Gesto Acción	11 - 15	11 - 20	500 - 1000	Genérico	Multivista	Interior
CAVIAR	RGB	Acción Interacción	6 - 10	Indefinido	< 20	Video-vigilancia	Multivista	Interior
HUMAN EVA I	RGB, MoCap	Gesto Acción	6 - 10	< 5	21 - 100	Genérico	Multivista	Interior
IXMAS	RGB, Siluetas	Gesto Acción	11 - 15	11 - 20	21 - 100	Genérico	Multivista	Interior
KTH Human Motion	RGB	Acción	6 - 10	> 21	500 - 1000	Genérico	Monovista	Exterior
WEIZMANN	RGB, Siluetas	Acción	6 - 10	6 - 10	21 - 100	Genérico	Monovista	Exterior

Como se puede observar en la mayoría de ellas, el método de grabación predominante es a través de cámaras RGB debido a la cantidad de información que proporcionan, a su bajo coste, su fácil utilización y su adaptabilidad frente a distintas situaciones y/o escenarios de grabación. Cabe destacar que aquellas bases de datos con mayor peso dentro de estos ámbitos, como son la KTH o la Weizmann, utilizan esta técnica.

Por otro lado, la mayoría de estas bases de datos estudian acciones genéricas, no existiendo demasiadas bases de datos específicas de video-vigilancia. Además, las bases de datos existentes recogen acciones muy concretas y en escenarios muy determinados, no acciones diarias como puede ser sentarse en una silla en un espacio público.

Se debe añadir que la gran mayoría de las bases de datos tratan de identificar y reconocer acciones en vez de simples gestos o actividades muy complejas.

Por estos motivos se va a proceder a realizar grabaciones de distintas secuencias estudiando acciones cotidianas que no han sido estudiadas antes, como por ejemplo sentarse o caerse al suelo, en un entorno de interiores a través de una técnica de vídeo. (Ver Capítulo 4)

Capítulo 3

Algoritmo de anotación del sistema

3.1. Introducción

Para poder desarrollar correctamente la interfaz de usuario, así como todo el proceso de anotación con sus diferentes opciones, se procede a explicar de una manera sencilla cómo se desarrolla el proceso de anotación.

El objetivo del algoritmo de anotación es la creación de un método sencillo para el usuario, a través del cual se pueda semi-automatizar el proceso de etiquetado en las secuencias de vídeo grabadas para crear la base de datos.

Este proceso de etiquetado consiste en marcar en las imágenes que forman el vídeo en qué lugar de la región de interés se encuentra cada usuario, llamando usuario a los actores participantes en las secuencias, qué usuario es y qué acción realiza de entre las estudiadas.

Para realizar el proceso de anotación partimos de una interfaz ya creada por el IDIAP, instituto de investigación afiliado a la Escuela Politécnica federal de Lausana, en Suiza ([2][17]), que se ha modificado de cara a cumplir los requisitos de este trabajo. La base de datos de IDIAP tiene como objetivo identificar la posición de las cabezas de dos personas y su seguimiento para realizar estudios acerca de la comunicación no verbal.

El algoritmo utilizado por este grupo de investigación realiza una extracción de los *frames* de los vídeos en formato JPG, y su posterior anotación para guardar los datos más relevantes.

Los datos se almacenan en ocho directorios, cada uno de ellos para cada secuencia. Cada directorio almacena los siguientes parámetros de interés:

- El número de la secuencia de vídeo
- La posición de la cabeza dentro del entorno
- Los parámetros intrínsecos de la cámara junto con los parámetros de calibración de la misma
- El número del *frame* del vídeo correspondiente
- Los ángulos de Euler de la orientación de la cabeza
- La localización 3D, tanto en píxeles como en centímetros, de la cabeza
- Un parámetro de confirmación de validez de los datos.

El objetivo de este proyecto, en cuanto a las anotaciones, es identificar la posición del cuerpo completo, no solo de la cabeza, almacenando las coordenadas "x" e "y" de la esquina superior izquierda y de la esquina superior derecha (ver Figura 34), así como la acción que se está llevando a cabo y el usuario que la realiza.

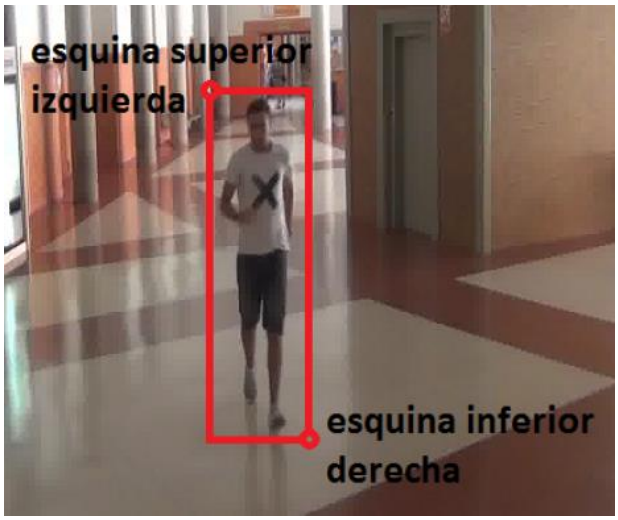


Figura 34 Coordenadas de interés para la anotación ([3])

Así mismo se podrá realizar el etiquetado cada cierto número de *frames,* de tal manera que el programa complete los *frames* que no han sido anotados manualmente. Esta información, si la anotación es manual o automática, también será almacenada junto con el número del *frame* correspondiente.

Para realizar el proceso de etiquetado y para obtener las secuencias de vídeo que forman la base de datos, se desarrolla una herramienta en MATLAB que sigue el diagrama general de la Figura 35.

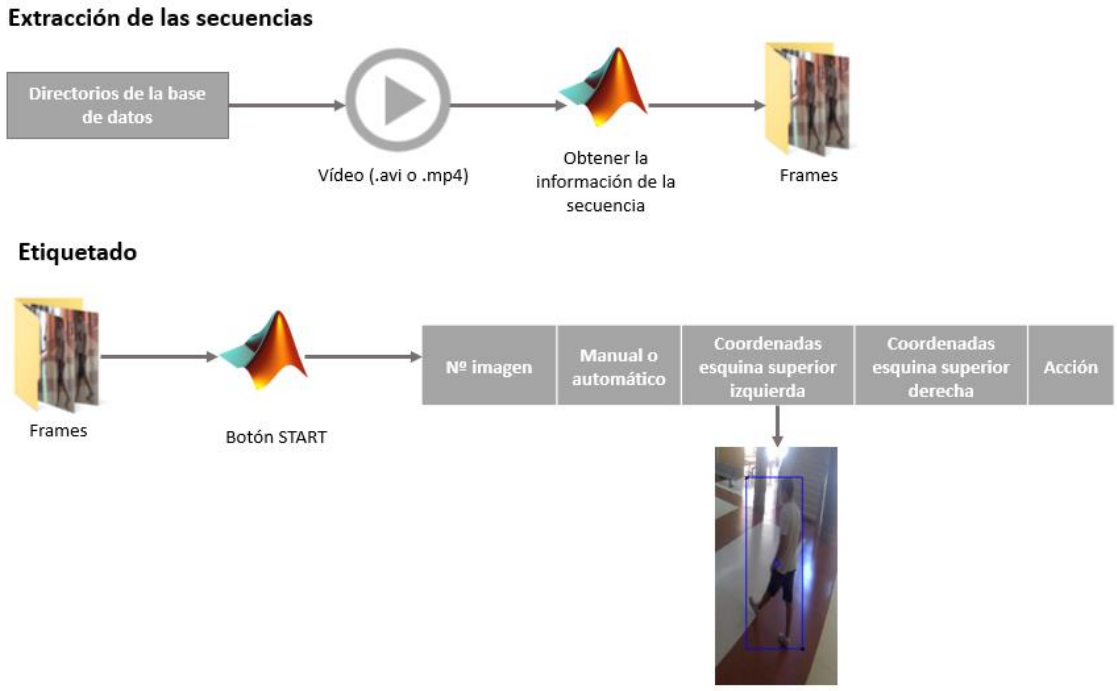


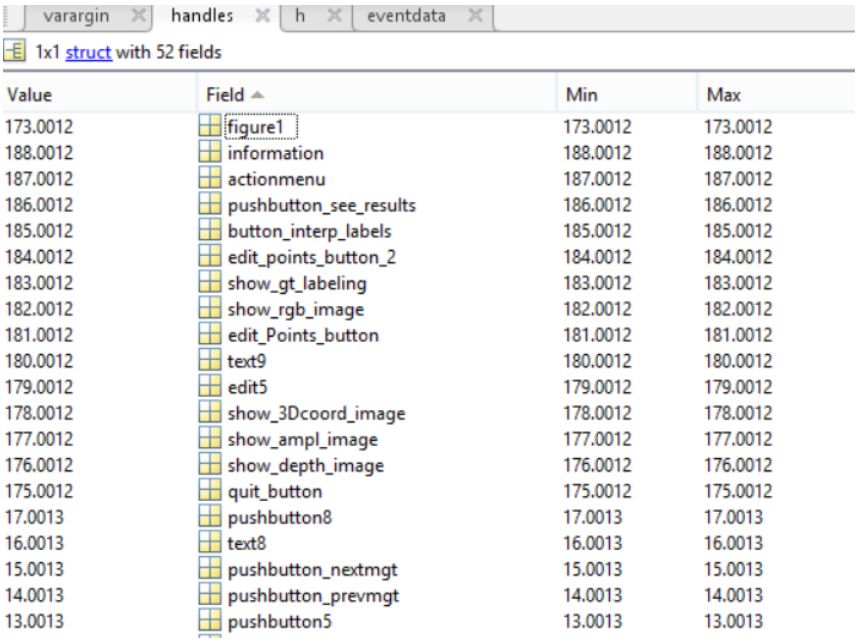
Figura 35 Diagrama general de anotación

El código realizado para este proyecto está basado en la interfaz gráfica de usuario en MATLAB, GUIDE, un entorno de programación visual para crear y ejecutar programas con necesidad de incluir datos dados por un usuario externo ([18]).

Este tipo de aplicaciones constan de dos archivos, un archivo .m, el cual contiene el código para controlar la interfaz, y un archivo .fig, el cual contiene los elementos gráficos para la interfaz.

GUIDE añade automáticamente sub-funciones al archivo .m asociadas a cada comando que se añada al archivo .fig, es decir, si añadimos un botón a nuestra interfaz, se generará de manera automática un *Callback* de dicho botón donde podremos desarrollar el código asociado a ese botón.

Además, se crea una estructura llamada *handles* donde se almacena todas las propiedades de los elementos de nuestra interfaz (tipo de botón, nombres de las variables, formato, etc) y los valores de las variables de todo el código. (La estructura de la Figura 36)


varargin handles h eventdata

1x1 struct with 52 fields

Value	Field	Min	Max
173.0012	figure1	173.0012	173.0012
188.0012	information	188.0012	188.0012
187.0012	actionmenu	187.0012	187.0012
186.0012	pushbutton_see_results	186.0012	186.0012
185.0012	button_interp_labels	185.0012	185.0012
184.0012	edit_points_button_2	184.0012	184.0012
183.0012	show_gt_labeling	183.0012	183.0012
182.0012	show_rgb_image	182.0012	182.0012
181.0012	edit_Points_button	181.0012	181.0012
180.0012	text9	180.0012	180.0012
179.0012	edit5	179.0012	179.0012
178.0012	show_3Dcoord_image	178.0012	178.0012
177.0012	show_ampl_image	177.0012	177.0012
176.0012	show_depth_image	176.0012	176.0012
175.0012	quit_button	175.0012	175.0012
17.0013	pushbutton8	17.0013	17.0013
16.0013	text8	16.0013	16.0013
15.0013	pushbutton_nextmgt	15.0013	15.0013
14.0013	pushbutton_prevmgt	14.0013	14.0013
13.0013	pushbutton5	13.0013	13.0013

Figura 36 Ejemplo de la estructura *handles* de la aplicación

En el código de esta aplicación se desarrollan distintas sub-rutinas para su correcto funcionamiento.

A continuación, se explicará cómo se desarrolla la fase de extracción de las secuencias y posteriormente se realizará lo mismo con la fase de anotación.

3.2. Extracción de secuencias

Para poder comenzar la anotación se deben tener imágenes que anotar. Debido a que se parte de vídeos grabados en formato .mp4, en primer lugar se deberán extraer todos los *frames* que forman el vídeo.

Al iniciar la interfaz, se procede a crear dos ventanas: una para la configuración y presentación de datos (que se va a explicar en detalle en este apartado) en la parte superior izquierda de la pantalla, y otra para la visualización de la imagen a etiquetar en la parte derecha (ver Figura 37).

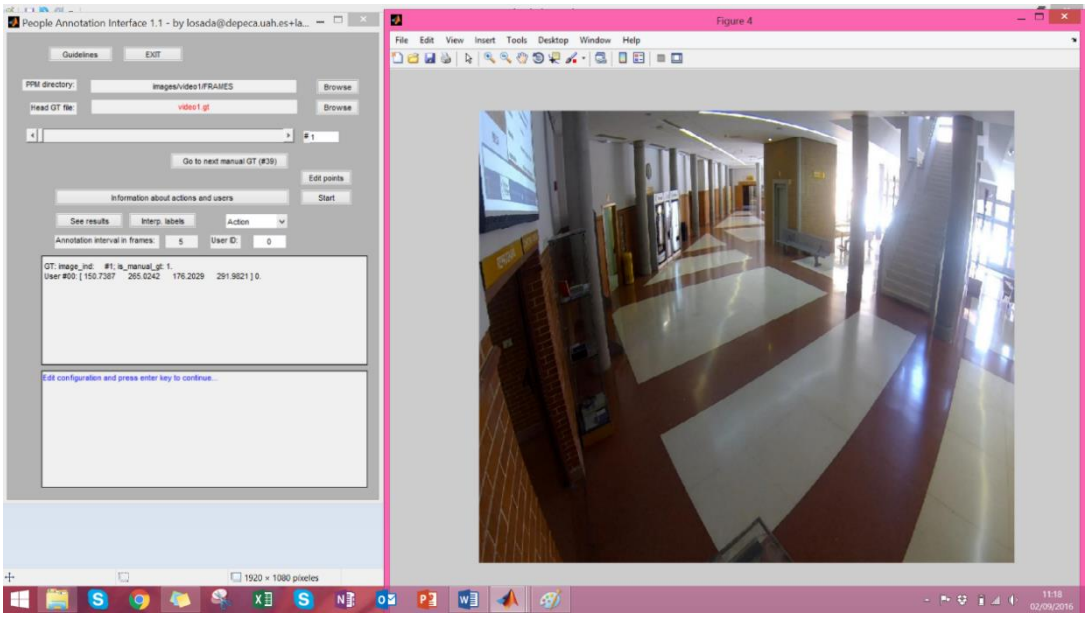


Figura 37 Figura de la interfaz (izq) y la figura de imágenes y anotación (dcha)

En la ventana de comandos se pueden encontrar distintas opciones:

- Comandos de elección de directorio
  Estos comandos, distribuidos en la parte superior de la ventana, se componen de dos botones de navegación por los distintos directorios y de dos barras con el nombre del directorio y archivo GT elegido. Estos comandos sirven para elegir el vídeo que se quiere anotar.

- Botones de navegación
  Este conjunto de botones está formado por un deslizador que sirve para navegar a lo largo de todos los *frames* del vídeo, así como para ver en qué punto del vídeo nos encontramos, por un cuadro donde aparece el número del *frame* que se está mostrando en la pantalla de anotación, y por dos botones para ir al siguiente, o al previo, *frame* etiquetado manualmente.

- Botones de control de anotación
  Son aquellos que sirven para realizar el proceso de etiquetado:
  - *START*
    Botón que sirve para comenzar a realizar el etiquetado
  - *EDIT POINTS*
    Botón para editar los puntos que se han etiquetado previamente
  - *INTERP. LABELS*
    Botón que sirve para completar el etiquetado manual y que sean anotados aquellos *frames* que no han sido etiquetados por el usuario mediante interpolación lineal.

- ♦ *SEE RESULTS*
  Botón que sirve para ver en la imagen de anotación los resultados del etiquetado de un vídeo

- Comandos para el establecimiento de parámetros
  Este conjunto de opciones sirve para que el usuario pueda elegir
  - El identificador del actor que aparece en el vídeo
  - La acción que realiza el actor
  - El intervalo que desea que haya entre un *frame* y otro en el proceso de anotación manual. Cuanto menor sea este intervalo, mayor precisión tendrán los resultados.

  En este proyecto las acciones elegidas para ser etiquetadas son:
  - Andar
  - Correr
  - Caerse
  - Sentarse
  - Quedarse quieto

- Opciones de información
  Para que esta interfaz sea de ayuda e intuitiva, se han diseñado una serie de elementos que sirven para que el usuario sepa cómo actuar:
  - Un botón *GUIDELINES*, que, al pulsarlo, te lanza un cuadro de diálogo donde explica brevemente la función de la aplicación
  - Un botón que sirve para obtener información sobre los usuarios, es decir, qué identificador tiene cada persona en el vídeo, y sobre acciones, descifrando el color al que está asociado cada acción
  - Dos ventanas blancas donde aparece información relativa al proceso de anotación. En la ventana superior la información que se proporciona es sobre los archivos GT, es decir, informa al usuario de si el *frame* seleccionado ha sido etiquetado o no. En la parte inferior aparece otra ventana blanca donde se puede ver un breve resumen del proceso cuando se abre el programa o bien unas breves indicaciones de qué teclas se deben utilizar para realizar el proceso de anotación

- Botón *EXIT*
  Botón para salir del programa de anotación

Una vez creadas las ventanas de la interfaz, se genera la estructura *handle,* donde se van a almacenar todas las variables y se inicializan aquellos parámetros considerados los más importantes del sistema:

- Figura a mostrar: dentro de las distintas posibilidades de este código de anotación se puede tener varios tipos de vista de la imagen a anotar, como por ejemplo imagen en 3D o imagen con profundidad. En este proyecto la figura a mostrar por defecto es la figura cuatro, la imagen en RGB
- Identificador de usuario 0
- Acción 0 (andar)
- Modo automático de anotación

En la estructura *handles* (ver Figura 36) podemos encontrar las variables principales de nuestro sistema:

- Los controladores de los botones de la interfaz de usuario
- Las opciones de la figura a mostrar. En este caso por defecto, la figura activa será la figura cuatro, que es la figura asociada a la imagen en RGB
- El identificador del usuario que se va a etiquetar
- Variable donde se almacena la acción realizada por el usuario
- El método de anotación: manual (1) o automático (0)
- El directorio de donde se deben extraer los *frames*
- El nombre de la secuencia que en la que se encuentra el programa, o la cual se va a proceder a anotar
- El listado de los nombres de los *frames* de una secuencia
- Los archivos GT de una secuencia. En cada celda del array se almacenan los datos que se obtienen de la anotación
- El número de *frame* de la secuencia

Se establecerán los parámetros por defecto, de tal manera que en la ventana de la interfaz únicamente se vean los botones asociados a los parámetros de la imagen en RGB, y se crean una serie de instrucciones para el usuario para que la interfaz sea más intuitiva.

El siguiente paso es buscar en la carpeta donde se encuentra el archivo .m algún directorio que contenga vídeos o *frames,* para inicializar el programa con el primero de ellos.

En caso de que no exista ninguna imagen extraída, se realiza la extracción de los *frames* para realizar la anotación. Así mismo se almacena la información necesaria para poder continuar con la aplicación, como es el directorio del vídeo seleccionado y el nombre del mismo para poder mostrarlo al usuario a través de la interfaz.

En el momento de extracción de los *frames* se reproduce una pantalla para que el usuario tenga conocimiento de cómo va el proceso de extracción.

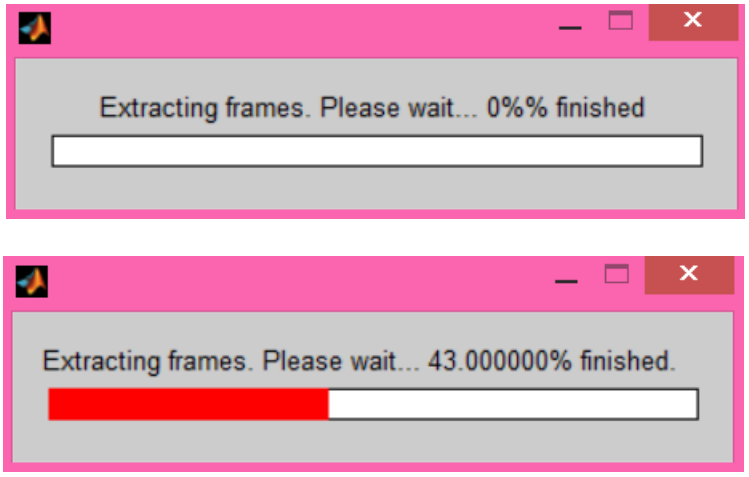


Figura 38 Visualización de la extracción de *frames*

Una vez que los *frames* se han extraído, o en el caso de que hayan sido extraídos anteriormente y estén almacenados en la carpeta FRAMES con formato .jpg, se procede a realizar una actualización de todos los parámetros que se ven afectados, a almacenar en la estructura principal todos los cambios realizados, así como reflejarlos a través de la interfaz.

Para entender mejor el proceso de extracción se procede a exponer un diagrama que lo clarifique (ver Figura 39)

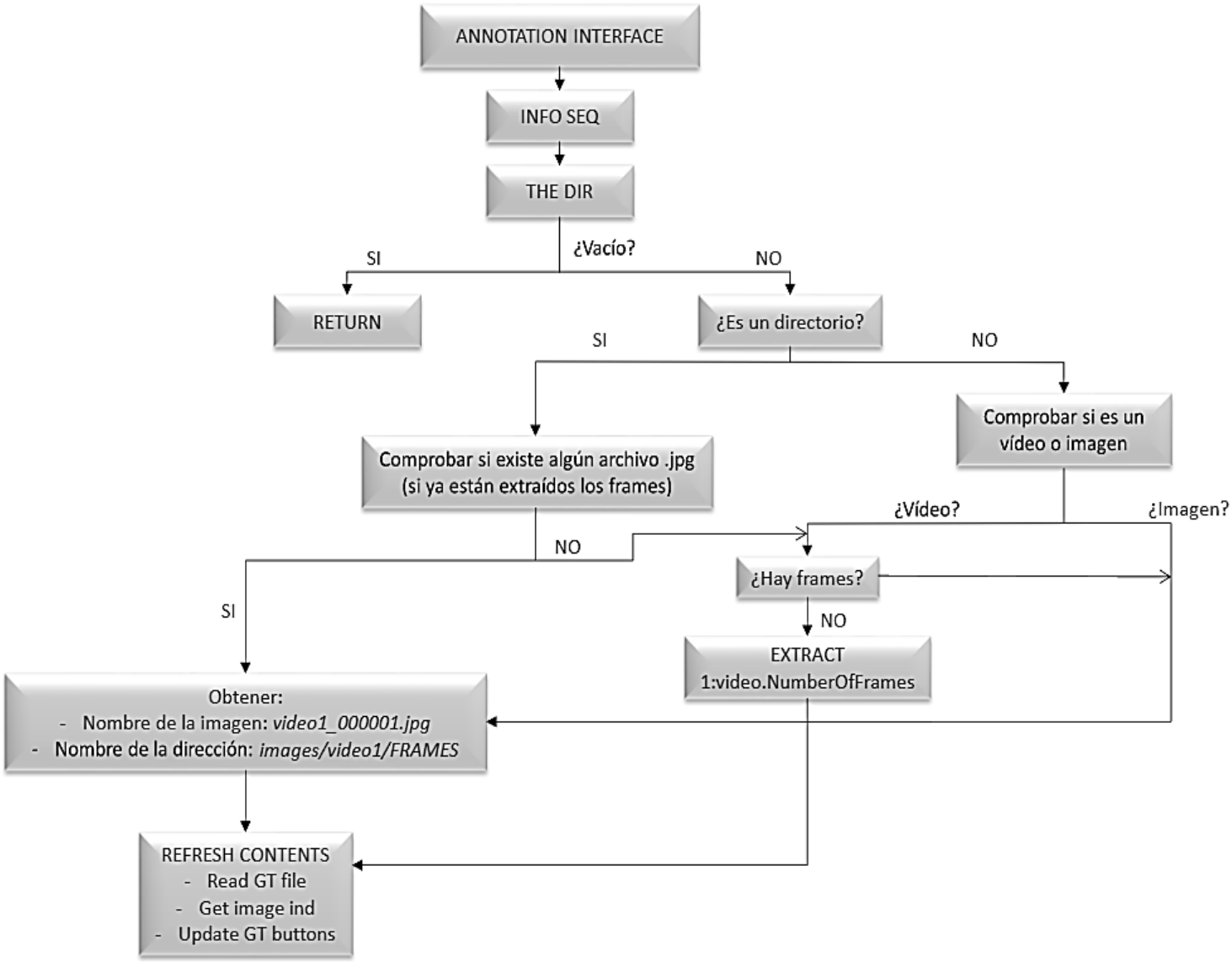


Figura 39 Diagrama de bloques de la fase de extracción de *frames*

Es importante añadir que solo es necesario realizar la extracción de los *frames* una vez por cada vídeo, ya que se quedan almacenados en la carpeta FRAME asociada a cada uno de ellos.

Una vez se ha realizado este proceso, se pasa a la segunda fase de la anotación, la fase de etiquetado.

3.3. Etiquetado de las secuencias

3.3.1. Elección del directorio

Una vez que se han creado las dos ventanas de la interfaz (ver Figura 37) se puede comenzar el proceso de etiquetado. Para ello, en primer lugar, se debe elegir en la interfaz de usuario, el vídeo que se desea anotar (Figura 40)

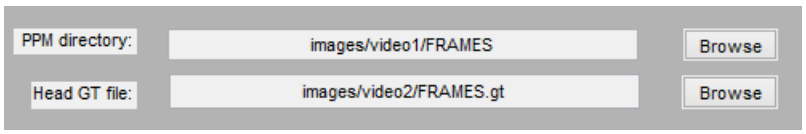


Figura 40 Elección del directorio

Para elegir el vídeo a etiquetar existen dos maneras de hacerlo.

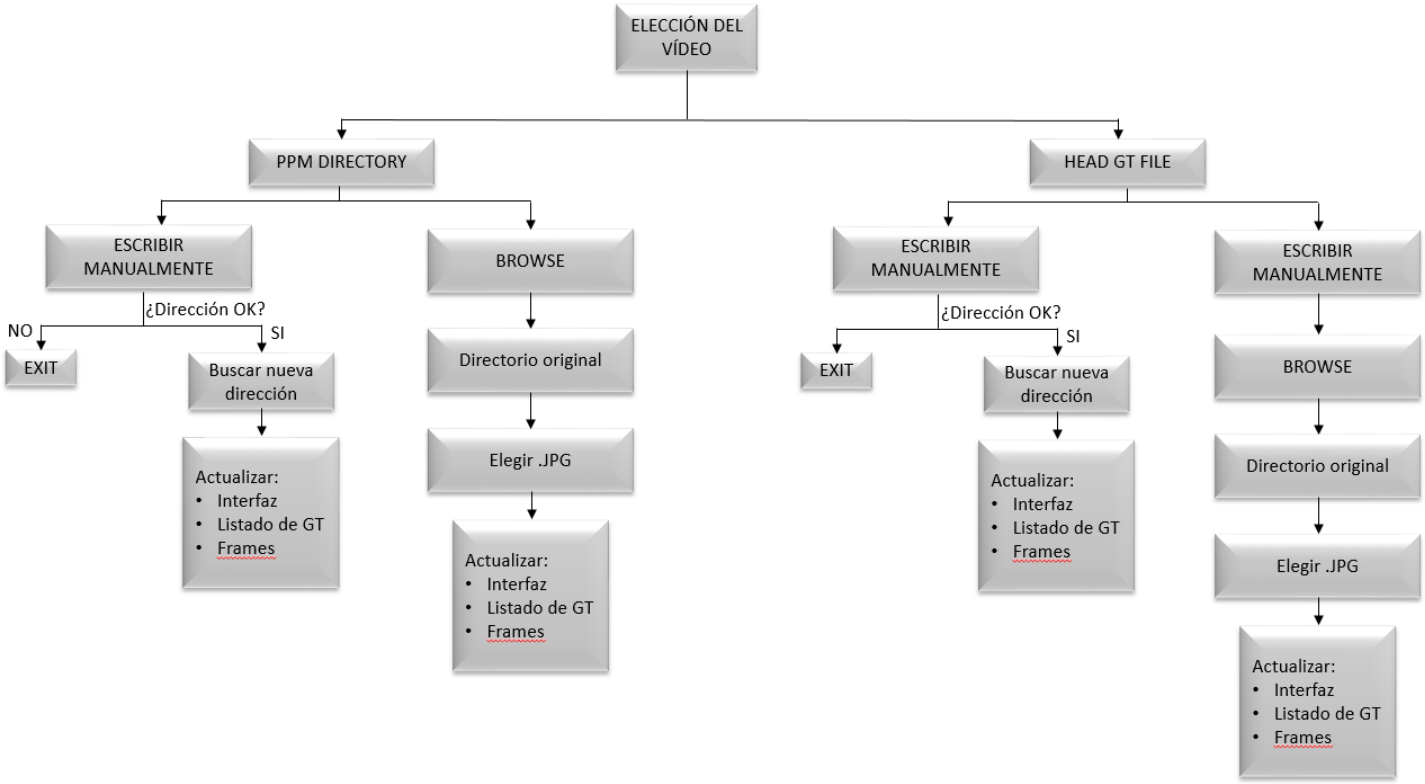


Figura 41 Elección del vídeo de trabajo

La primera opción es elegir el directorio donde se encuentra la carpeta FRAMES del vídeo elegido[1] (PPM Directory). Para ello existen dos posibilidades: escribir manualmente el directorio, o elegirlo pulsando el botón BROWSE de la parte superior de la Figura 40.

[1] En el caso de que no exista la carpeta FRAMES de dicho vídeo se procederá a su extracción (3.2)

1. Elección del directorio manualmente

Cuando se escribe manualmente el directorio, el programa accederá a la pestaña donde está la nueva dirección escrita.

En el caso de que el directorio escrito sea correcto, se realiza búsqueda de la nueva dirección y se obtiene el acceso a la carpeta correspondiente.

Posteriormente se realiza una actualización de todas las variables: el listado de archivos GT de la nueva secuencia, los *frames*, el número de *frames* extraídos, los archivos GT manuales, y, por ende, la actualización de los botones *got to the next/previuous manual GT,* etc.

2. Elección del directorio a través del botón BROWSE

Al hacer click en el botón *BROWSE* el programa accederá a su sub-rutina correspondiente. En primer lugar, abrirá el directorio original el cual se abre por defecto, para que así el usuario pueda navegar por las carpetas que desee y elija la carpeta de FRAMES que quiera abrir.

El usuario deberá elegir el *frame* a abrir, es decir, el archivo .jpg. Una vez elegido, el programa almacenará la nueva dirección seleccionada y descargará, al igual que en los procesos anteriores, toda la información necesaria para poder abrir la nueva secuencia.

La segunda opción para elegir el vídeo a etiquetar es seleccionando el archivo GT. Para realizarlo, como se ha dicho anteriormente, existe la opción de escribir manualmente el nombre del archivo (con el directorio incluido), o bien seleccionarlo en el directorio correspondiente a través del segundo botón *BROWSE* de la Figura 40.

1. Elección del archivo GT manualmente

Cuando se escribe manualmente el nombre del archivo, el programa, en primer lugar, procede a leer lo escrito por el usuario y comprobar que el directorio es el correcto. En el caso de que no lo sea se imprimirá por pantalla un mensaje de error:

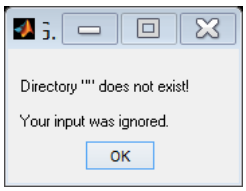


Figura 42 Mensaje de error de directorio

En el caso de escribir correctamente la dirección (la manera correcta es la mostrada en la Figura 40) se obtiene tanto el nuevo *path* como todos los datos necesarios para cambiar el directorio y actualizar la imagen de la figura de etiquetado.

2. Elección del archivo GT a través del botón BROWSE

Cuando se pulse el botón BROWSE el programa procederá a abrir el *path* original para que el usuario elija el nuevo directorio en el que desee trabajar. El proceso es el mismo que los casos anteriores.

3.3.2. Proceso de anotación

Una vez que se ha elegido el vídeo que se desea etiquetar, y que se han abierto y descargado los *frames* que lo componen, se puede comenzar a realizar la anotación.

Para ello el usuario deberá hacer click en el botón START de la interfaz de usuario (ver Figura 43). Primeramente, para hacerle al usuario más fácil el proceso, se mostrará la información necesaria por pantalla para realizar el etiquetado (ver Figura 44).

Figura 43 START

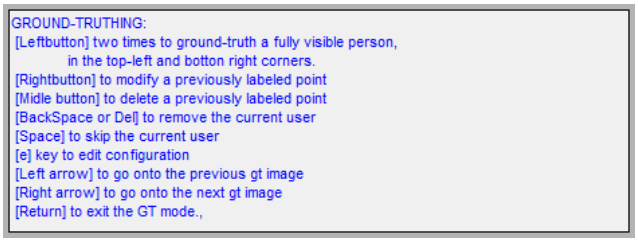
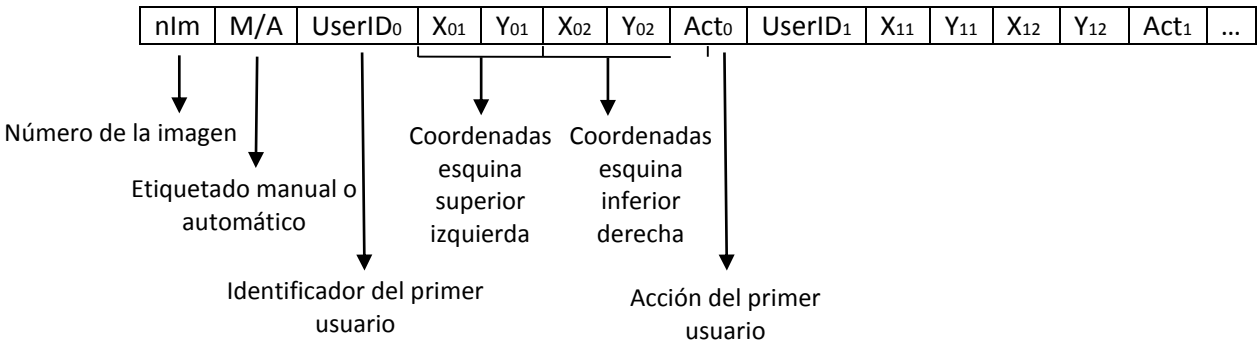


Figura 44 Información para anotación

Posteriormente se leerán las variables necesarias para crear el archivo de *ground truth* correspondiente. En esos archivos GT se almacena una línea por cada imagen etiquetada, la cual contiene la siguiente información:

- El número de la imagen dentro de la secuencia
- Una variable que indica si el etiquetado es manual (1) o automático (0)
- Por cada usuario etiquetado en la imagen aparecerán 6 valores:
  - El identificador del usuario
  - Las cuatro coordenadas “x” e “y” de la esquina superior izquierda y la esquina inferior derecha
  - El número asociado a la acción realizada

Tabla 8 Almacenamiento del archivo GT



Primero se obtendrá, a través de la interfaz, el identificador del usuario que se desea etiquetar y la acción a realizar. Por otro lado se obtendrá el número de la imagen en la que nos encontramos y número de *frames* extraídos de la secuencia.

Una vez que el programa disponga de estos datos, mostrará en la figura de anotación unos ejes para obtener las coordenadas del etiquetado:

Figura 45 Ejes de anotación

Cuando se han marcado las dos esquinas correspondientes, se almacenan las coordenadas en píxeles como se muestra en la Tabla 8

En el caso de que no exista un etiquetado previo en el *frame* donde nos encontramos, se almacena en primer lugar el número de la imagen, a continuación, el modo de etiquetado (manual o automático), seguido del identificador de usuario y las coordenadas de este, y por último la acción:

```
000002 1 0000   352.9267    83.8386   506.5267   316.2795 0
```

En este ejemplo se observa el archivo GT generado tras etiquetar un *frame* que estaba vacío. Primero se almacena el número de la imagen, en este caso el 2, y al ser un etiquetado manual, el siguiente dato a almacenar es un 1. En la tercera posición del array se ve como el usuario tiene el identificador 0 y las cuatro coordenadas almacenadas. Por último, se observa que la acción guardada es la identificada por el número 0.

En el caso de que sí exista un etiquetado previo, ya que por cada *frame* solo puede haber una línea en el archivo GT, el programa almacenará la información a continuación de los datos ya guardados, pero sin incluir el número de la imagen ni el modo de etiquetado.

```
00071 1 0001 96.98 25.28 12.95 68.48 0 0004 58.72 18.64 68.69 44.86 0
```

En este ejemplo se puede ver cómo, tras realizar el etiquetado de un *frame* donde el usuario 1 ya estaba etiquetado, los datos del nuevo usuario, en este caso el 4, son almacenados a continuación de los datos previos. Los datos marcados en azul son los que corresponden al usuario 1 y los datos marcados en color rojo son los asociados al usuario 4.

Como se puede comprobar, el número de imagen y que la anotación es manual son elementos compartidos por ambos usuarios.

Una vez que se obtienen todos los datos del usuario y estos son almacenados, se procede a mostrar en la interfaz el cuadro de anotación obtenido.

Durante el proceso de anotación, el cuadro de etiquetado del usuario que está siendo anotado se marcará de color blanco. En el caso de que existan otros usuarios en el mismo *frame*, los usuarios que no estén siendo anotados tendrán su caja de color negro.

Figura 46 Caja de etiquetado

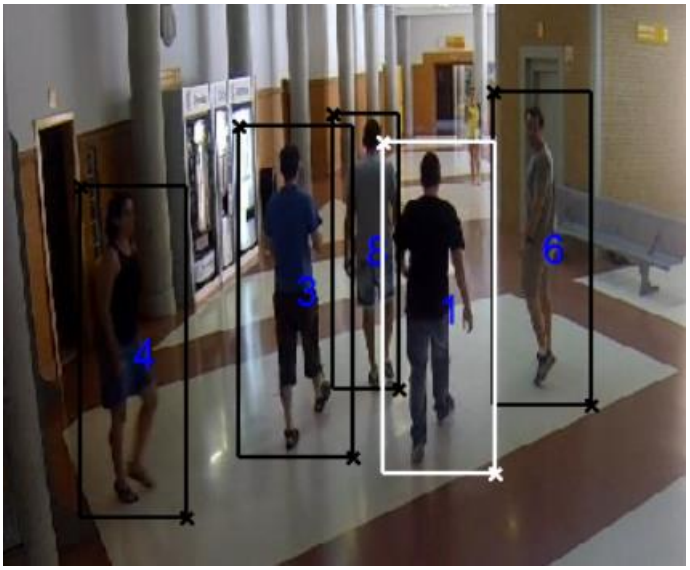


Figura 47 Cajas de etiquetado múltiples

En el caso de etiquetar un *frame* que ya tiene datos almacenados, el programa comprueba que el usuario que estamos etiquetando no ha sido anotado previamente. En el caso de que no haya sido etiquetado se coloca el identificador del nuevo usuario, sus coordenadas y la acción a continuación de los datos almacenados anteriormente, como ya se ha visto.

Si el usuario ha sido etiquetado previamente, se procede a modificar los datos anteriores por los nuevos. Una vez realizado este proceso se actualiza la imagen y se procede a almacenar todos los datos obtenidos.

Para entender mejor el proceso de anotación se muestra el diagrama pertinente.

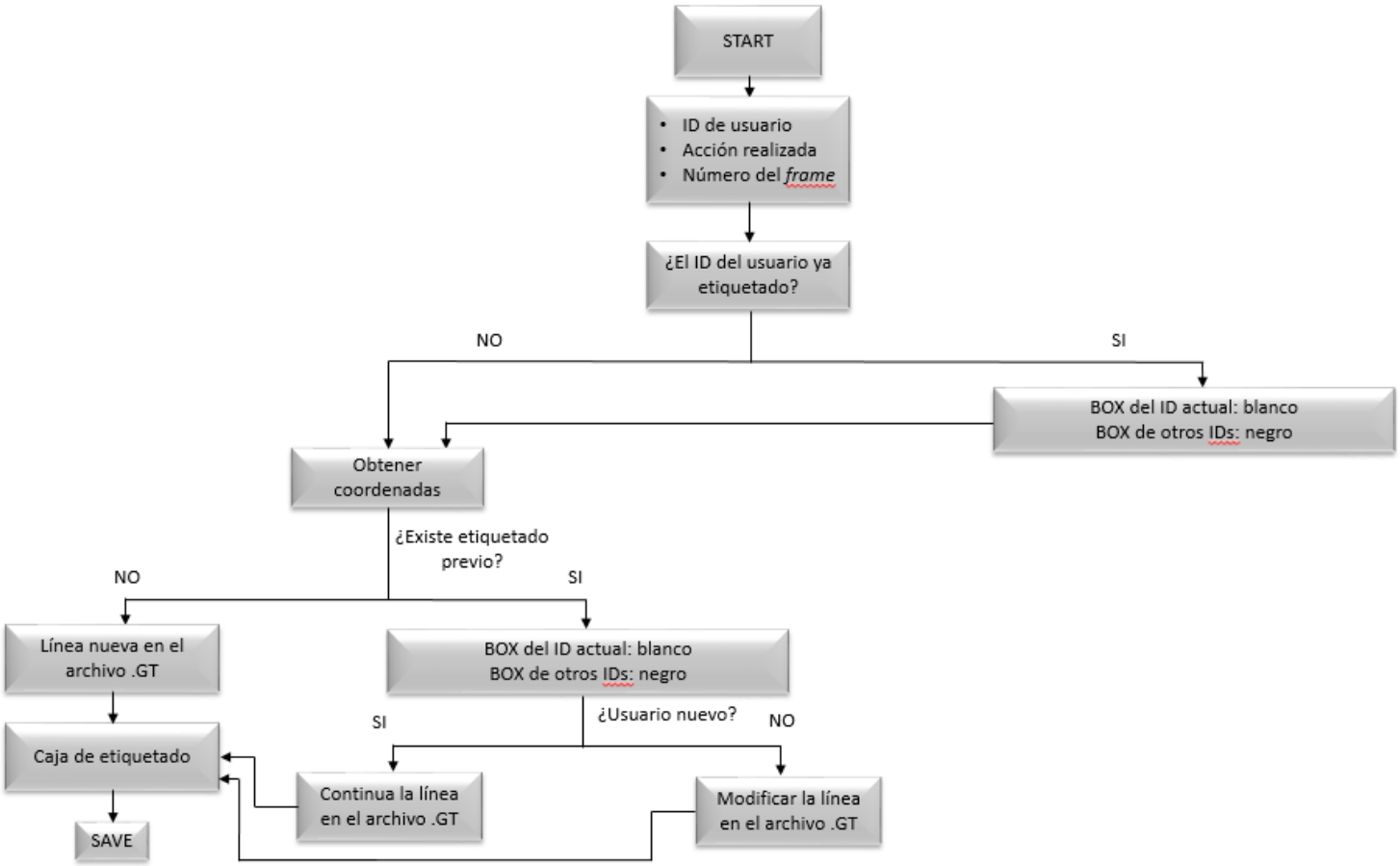


Figura 48 Diagrama del proceso de anotación

3.3.3. Opciones durante proceso de anotación

El programa por defecto vuelve a mostrar los ejes de etiquetado para: editar los puntos que se han marcado como se ha visto anteriormente, etiquetar un nuevo usuario en la imagen, etiquetar el siguiente *frame,* editar la configuración, o finalizar el proceso de anotación.

Para ello lee el código ASCII que corresponde a cada tecla, como se puede ver en la Tabla 9

Tabla 9 Código ASCII de las teclas y sus funciones

CÓDIGO ASCII	TECLA	ACCIÓN A REALIZAR
28	Flecha izquierda	Pasar al anterior *frame* para etiquetar
29	Flecha derecha	Pasar al siguiente *frame* para etiquetar
101	e	Editar la configuración
32	Barra espaciadora	Salir del modo etiquetado
127	Suprimir/Delete	Eliminar el usuario actual
27	Escape	Salir del modo anotación

1. *Pasar al siguiente frame para etiquetar*

Al pulsar la tecla derecha de las flechas, el programa leerá el código ASCII correspondiente y se irá a la parte del código asociada con dicho código. Primeramente se obtiene el número del *frame* donde está el etiquetado actual, y posteriormente se obtiene el intervalo de anotación que haya elegido el usuario. Estos valores se suman y, si son menores que el máximo de *frames* que tiene el vídeo (para no excederse), se procede a realizar las actualizaciones oportunas:

- Se actualiza la barra de SLIDER
- Se actualiza la imagen que se ve en la figura de anotación, pasándole para ello el nuevo número de *frame*
- Actualiza el botón *go to the next manual GT*
- Actualiza el botón *go to the previous manual GT*
- Actualiza los datos mostrados a través de la interfaz

2. *Pasar al anterior frame para etiquetar*

Si se pulsa la tecla izquierda se procederá a ir al *frame* anterior que ha sido etiquetado manualmente. Una vez leído el carácter ASCII asociado a esta tecla, se obtiene el intervalo de anotación establecido por el usuario y se le resta al *frame* en el que el programa se encuentra. Posteriormente se procede a realizar las mismas actualizaciones que en el sub-apartado anterior

3. *Editar la configuración*

Para editar la configuración mientras el usuario se encuentra en el proceso de etiquetado, se debe pulsar la letra e. Al hacerlo, el programa acude a la parte del código asociado y procede a mostrar por pantalla un mensaje donde se indica que se debe modificar la configuración:

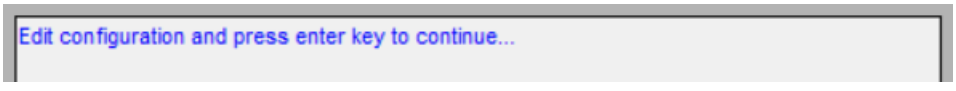


Figura 49 Editar la configuración

El programa esperará 30 segundos a que el usuario realice las modificaciones oportunas y continuará ejecutándose con normalidad para continuar etiquetando.

4. Salir del modo etiquetado

En el caso de que el usuario presione la barra espaciadora el programa se mantendrá en el mismo *frame* pero saldrá del modo etiquetado, es decir, saldrá de la función *START*, pero no del programa de anotación.

5. Eliminar el usuario actual

Si se presiona la tecla *suprimir* del teclado, primero se buscará en el archivo GT la línea correspondiente al número de imagen en el que esté en ese momento el programa. Después, se buscará en dicha línea el usuario elegido (el que aparece escrito en la interfaz) y posteriormente se procede a eliminar del array las coordenadas, la acción y el identificador del usuario correspondiente.

6. Salir del modo anotación

En el caso de presionar la tecla ESC se procede a realizar un guardado de todos los archivos GT y los etiquetados creados previamente y se sale del programa mostrando por la pantalla de comandos el siguiente mensaje:

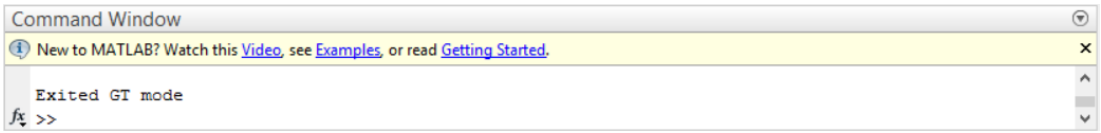


Figura 50 Comando ESCAPE

3.3.4. Fin del proceso de anotación

Debido a que en la interfaz de usuario existe un *slider* para mostrar el progreso de etiquetado, se puede comprobar cuándo se ha acabado de etiquetar y finalizar presionando la tecla de escape.

En el caso de que el usuario no se dé cuenta de ello, si en el proceso de anotación el usuario presiona la flecha derecha del teclado para ir al siguiente *frame,* y no se puede avanzar debido a que se ha acabado la anotación, el programa mostrará un mensaje informativo (ver Figura 51) y saldrá del modo de anotación.

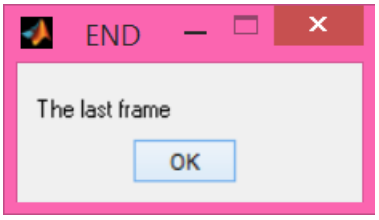


Figura 51 END de la anotación

Una vez que se ha finalizado la anotación de un vídeo, existen varias opciones a tener en cuenta. Es posible que el usuario en algún *frame* cometa algún tipo de error en la anotación, o que simplemente se quiera realizar un etiquetado más preciso de algún usuario.

Además, el usuario querrá ver los resultados de la anotación. Por último, se deberá realizar una interpolación lineal de los puntos anotados manualmente para completar la anotación de todos los *frames*.

3.3.4.1. Edición de puntos

Para editar las coordenadas de un usuario existe la opción de editar los puntos de la caja de etiquetado a través del botón *EDIT POINTS*.

Figura 52 Botón de edición de puntos

Esta función actualiza la imagen de anotación, es decir, se pone en modo anotación, siendo los colores de las cajas de etiquetado negro para usuarios ajenos y blanco para el usuario seleccionado.

Una vez en modo anotación, se obtienen los parámetros de ese *frame*, como el número de imagen, y realiza una búsqueda a través del archivo GT para comprobar si ese número de imagen está etiquetado. En el caso de que no sea así, el programa mostrará un cuadro de diálogo informando sobre ello.

Si el *frame* en cuestión sí está etiquetado, el programa realizará una búsqueda del ID de usuario en el número de imagen correspondiente, de sus coordenadas y de la acción realizada por el mismo. Una vez almacenada toda esta información en una variable auxiliar, se mostrará por pantalla los ejes de la Figura 45 para tomar las nuevas coordenadas.

Cuando se hace el primer click, el que en teoría corresponde con la esquina superior izquierda, el programa realiza una búsqueda del punto más cercano que haya sido etiquetado, y lo modifica poniendo el nuevo valor, eliminando los puntos antiguos e incluyendo los nuevos. Vuelve a realizar lo mismo con el segundo punto, el que correspondería con la esquina inferior derecha.

Una vez que los puntos han sido editados se puede proceder a salir del modo edición de puntos presionando la tecla ENTER.

3.3.4.2. Visualización de los resultados

Por último, para ver los resultados de la anotación, existen dos opciones.

En primer lugar, los resultados pueden ser vistos si se va deslizando el *slider* de la interfaz de usuario. Ahí se podrán ver los recuadros de cada actor con su color determinado.

Esto quiere decir que, durante el proceso de anotación, los colores de las cajas de etiquetado eran blanco para el usuario que se estaba etiquetando y negro para los usuarios previamente etiquetados.

Cuando se sale del proceso de anotación, esta interfaz modifica esos colores en función de la acción que esté desarrollando el actor:

Tabla 10 Colores asociados a cada acción

COLOR	ACCIÓN
Azul	Andar
Rojo	Correr
Amarillo	Sentarse
Verde	Caerse
Rosa	Estar en el sitio

Por ejemplo, en la figura Figura 53 los usuarios 0, 1, 6, 7 y 8 tienen el cuadro de etiquetado de color rosa, lo que implica que no están realizando ninguna acción, están quietos en el sitio.

El usuario con el identificado 2 está andando ya que tiene la caja de color azul, los usuarios 3 y 4 están sentándose, ya que su color es el amarillo, y por último, el usuario de identificador 5 está corriendo por el pasillo.

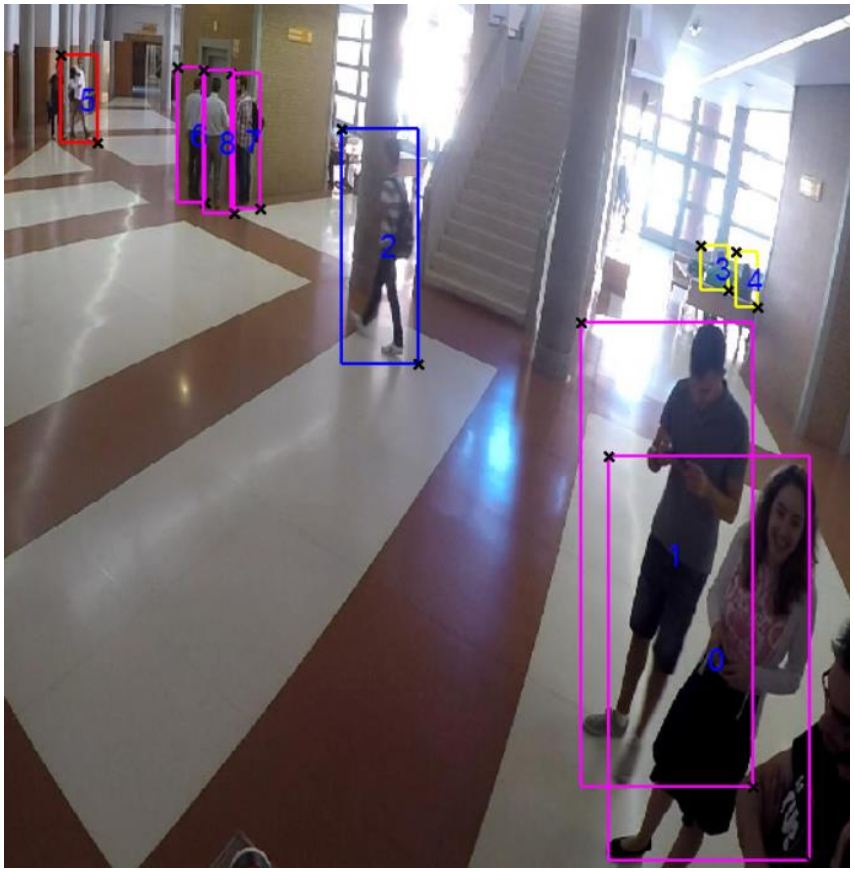


Figura 53 Etiquetado de varias acciones

En segundo lugar, otra de las opciones a la hora de visualizar los resultados es a través del botón *SEE RESULTS*. Al presionar dicho botón se muestra por pantalla las indicaciones necesarias para poder ver la secuencia completa y cómo detenerla o trabajar con ella:

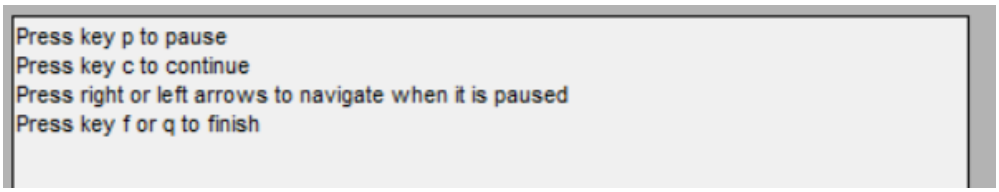


Figura 54 Directrices del botón *see results*

Una vez mostradas las directrices, se descargará el número de imágenes de la lista de *frames* asociada al vídeo, y desde la imagen actual hasta el total de número de imágenes, a través de un bucle, se irá mostrando en la pantalla de anotación, todos los *frames* en orden, marcando dónde está cada usuario.

En este caso, en vez de marcar la caja de etiquetado, el programa marca el punto medio de dicha caja con el mismo código de colores anterior. El motivo por el cual en esta opción solo se ve el punto y no la caja, es para que la imagen no se vea demasiado cargada.

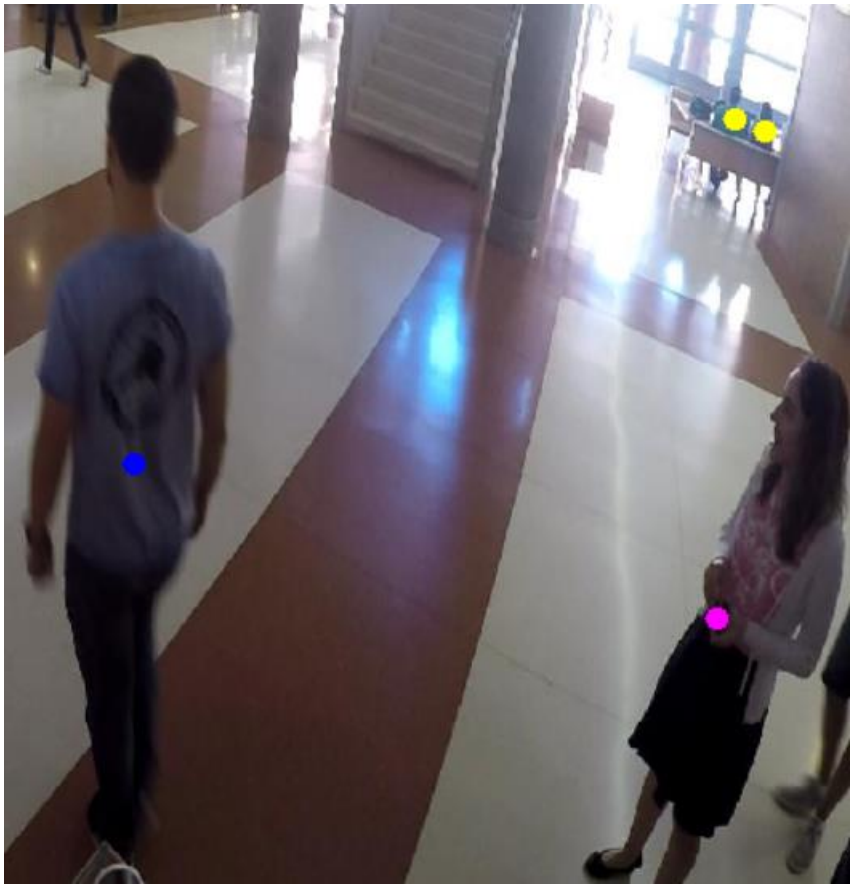


Figura 55 Imagen mostrada por el botón *see results*

3.3.4.3. Interpolación de los puntos de anotación

Para completar el proceso de anotación, deben ser etiquetados aquellos *frames* que no han sido anotados manualmente. Para ello existe un botón en la interfaz, llamado *Interp. Label*, que realiza una interpolación lineal entre los puntos ya etiquetados.

El diagrama principal de este proceso se puede ver en la Figura 56.

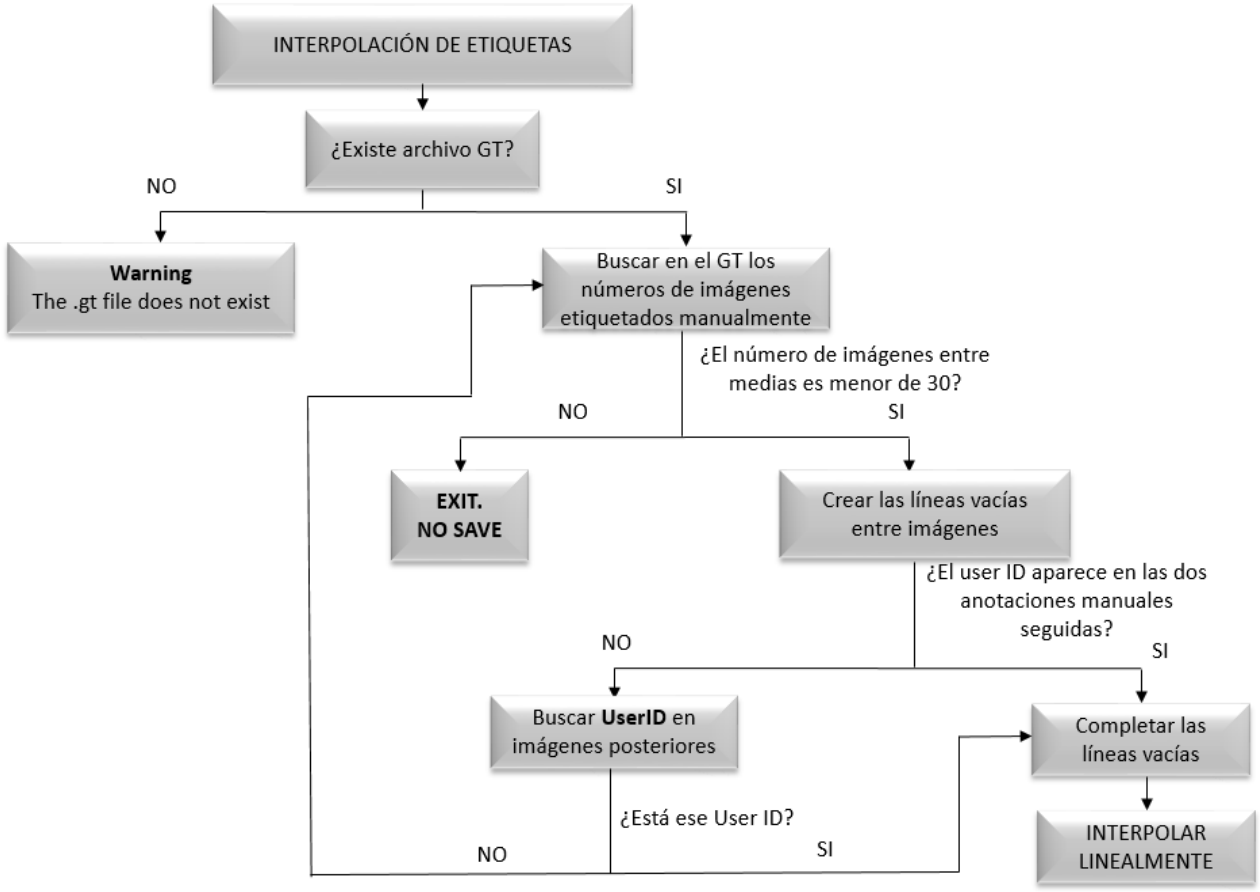


Figura 56 Diagrama principal de interpolación

Para realizar todo este proceso, primeramente, se extrae todo el listado de GT manuales del vídeo en cuestión (ver Figura 57) y se comprueba qué imágenes son las que están anotadas manualmente, creando para ello un listado de los números de dichas imágenes (ver Figura 58).

1	000061	1	0000	916.3621	239.1895	1087.8210	617.7231 1
2	000066	1	0000	902.7812	238.0663	1019.9164	570.5468 1
3	000071	1	0000	860.3408	215.6014	972.3833	553.6981 1
4	000076	1	0000	819.5981	206.6154	953.7095	526.7402 1
5	000081	1	0000	802.6220	202.1225	933.3382	471.7012 1
6	000086	1	0000	783.9483	194.2598	880.7122	459.3456 1
7	000091	1	0000	763.5769	181.9041	855.2480	450.3596 1
8	000096	1	0000	746.6008	177.4111	836.5743	413.2925 1

Figura 57 Ejemplo de archivo GT

	1	2
1	61	
2	66	
3	71	
4	76	
5	81	
6	86	
7	91	
8	96	
9		

Figura 58 Listado de los números de imágenes anotados manualmente

Una vez obtenidos estos datos se va comprobando el número de imágenes que hay entre los *frames* anotados manualmente. Es decir, en este ejemplo, entre la imagen 61 y la 66, existen 4 *frames* que se deberían de rellenar.

En el caso de que el número de imágenes sea menor que 30, que es el número máximo permitido, se procede a realizar la interpolación. Esos 30 *frames* corresponden con casi medio segundo de grabación y se elige para tener la mayor precisión posible en cuanto al etiquetado se refiere.

Inicialmente, se obtiene el ID de usuario y se comprueba si este usuario ha sido etiquetado en el *frame* previo.

En el caso de que el usuario sí haya sido etiquetado previamente, se crean las nuevas líneas de las imágenes no anotadas, en este caso, se crean las líneas 62, 63, 64 y 65, y se coloca en la siguiente celda de cada una de las líneas un 0, que representa el etiquetado automático. Posteriormente se añade el ID de usuario.

A continuación, se procede a calcular el incremento de distancias. Para ello se realiza la siguiente operación:

$$\frac{Coordenadas\ de\ la\ primera\ imagen\ manual - Coordenadas\ de\ la\ segunda\ imagen\ manual}{Número\ de\ imágenes\ entre\ las\ imágenes\ manuales - 1}$$

Con el resultado obtenido se calcula el incremento que debe existir entre cada una de las imágenes para llegar de la coordenada de la primera imagen manual, en nuestro ejemplo la 61, hasta la segunda imagen manual, en nuestro ejemplo la 66.

Para entenderlo mejor se explica con el ejemplo. Primero se selecciona la primera coordenada, 916,3621. A este valor se le resta la coordenada de la siguiente imagen manual, es decir, se le restaría 902,7812. El resultado obtenido, 13,5809, debe ser dividido entre el número de imágenes que hay entre ambos puntos iniciales, que serían 5 imágenes. El resultado de la operación final es 2,71618.

Este valor último es el incremento de cada coordenada de imagen en imagen:

62	0	0	913.6459	238.9649	1.0742e+03	608.2878
63	0	0	910.9297	238.7402	1.0607e+03	598.8526
64	0	0	908.2136	238.5156	1.0471e+03	589.4173
65	0	0	905.4974	238.2909	1.0335e+03	579.9821

Figura 59 Interpolación de las coordenadas

Por último, se copia la acción desarrollada por el usuario y se rellena con toda esta información el archivo GT original, para así finalizar el proceso de automatización de la anotación.

Este proceso ocurre en el caso de que el ID de usuario sea el mismo en ambos *frames* manuales.

Si entre imágenes anotadas manualmente, aparece un ID nuevo, el programa obtiene el nuevo ID y lo busca en las siguientes anotaciones manuales, y realiza el mismo proceso que se ha contado antes.

Por último, si el número de imágenes entre anotaciones manuales es mayor que 30, el número máximo de interpolación, el programa se sale de este modo y no guarda nada del proceso de interpolación.

3.4. Conclusión del desarrollo de la interfaz

Como se ha podido comprobar, la interfaz de usuario es lo más intuitiva posible. El proceso de creación del algoritmo de anotación ha sido un proceso largo que ha pasado por distintas personas de distintas partes del mundo, y que puede mejorarse de diferentes maneras.

En este proyecto se ha incluido todo lo relacionado con la parte de la acción, tanto en la parte de la interfaz visible, como en el código, y se han realizado ciertas mejoras en la interfaz de usuario para hacerla algo más amigable, como por ejemplo la inclusión de información sobre las acciones y los usuarios.

Para ello se ha tenido que entender el código y el desarrollo de todo el proceso, ampliando los archivos GT, modificando las cajas de anotación y sus colores de identificación, cambiando ciertos parámetros de la interpolación de etiquetas, así como la edición de puntos, y realizando los cambios oportunos en todo el código afectado.

Capítulo 4

Evaluación

Una vez desarrollado el código y realizados los vídeos de interés, se procede a mostrar los resultados obtenidos, es decir, la base de datos que ha sido implementada.

4.1. Resultados

La base de datos creada está formada por diecinueve vídeos de distintos tamaños en los cuales se estudian cinco acciones en un entorno de interiores. Estas acciones están realizadas por diferentes actores, tanto hombres como mujeres, de manera natural, es decir, no han sido ensayadas previamente.

4.1.1. Escena de grabación

Todos los vídeos han sido grabados en el interior de la Escuela Politécnica de la Universidad de Alcalá a través de una cámara GoPro HERO4.

Esta cámara es sencilla, práctica y liviana, lo que hace que sea fácil grabar con ella y adaptarla a distintos entornos. Tiene una resolución de 1280x720 píxeles, captura imágenes a 50fps y tiene un campo de visión ultra gran angular, siendo el ángulo de visión mayor al de la visión humana (los ángulos de visión de este tipo de cámaras oscilan entre los 60º y 180º). El formato de almacenamiento de los vídeos es mp4. ([19])

Para realizar los vídeos la cámara está colocada en la zona sur de la Escuela, enfocada con orientación noroeste y en una zona con cierta altura, para así grabar la máxima región y reducir las posibles oclusiones.

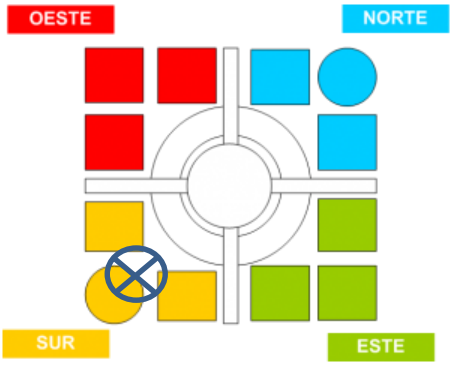


Figura 60 Localización de la cámara en la EPS

El la Figura 61 se muestra el área grabada, así como la región de interés, la cual va a ser etiquetada.

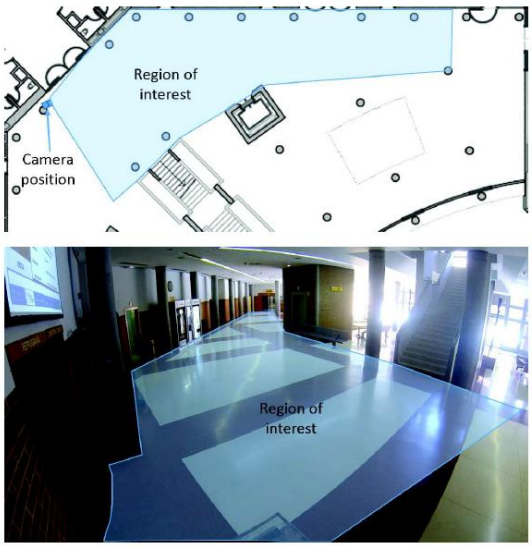


Figura 61 Región de interés de la base de datos

Como se puede observar, la región de interés es un pasillo formado por unas escaleras a la derecha situadas entre dos columnas, un ascensor y algunos bancos, los cuales serán movidos y recolocados en función del objetivo del vídeo.

Además, esta es una de las zonas más concurridas de toda la Escuela, lo que hace que aparezcan en escena personas que no son los actores "oficiales" de los vídeos, llamados "intrusos". Así mismo, la iluminación de la región de interés depende de la hora del día y no está controlada, puesto que es iluminación natural.

Debido a todo esto, esta base de datos es totalmente realista y natural, justo lo que se necesita para realizar video-vigilancia en interiores. ([20])

4.1.2. Acciones estudiadas

Las acciones elegidas para el estudio son aquellas que se han considerado las más interesantes para la aplicación a la que va dedicada esta base de datos, que es video-vigilancia en espacios cerrados. Estas acciones son andar, correr, sentarse y caerse.

Además, en la aplicación se puede etiquetar también a aquellas personas que están estáticas en la imagen sin realizan acción alguna.

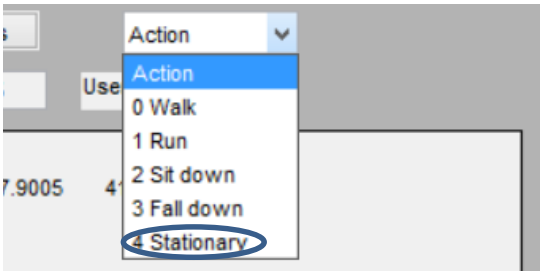


Figura 62 Listado de acciones y opción *stationary*

Para la realización de estas acciones se eligen a 16 personas, llamadas "usuarios", de los cuales dos son mujeres y el resto hombres. Cada uno de estos usuarios tiene su propio

identificador, y aquellas personas que aparecen en las imágenes que no son usuarios, si no intrusos, serán etiquetadas por orden de aparición desde el número 100 en adelante.

Cuando han sido anotados, el identificador del usuario se podrá ver en el punto medio de la caja de etiquetado, y las cajas variarán su color dependiendo de la acción que esté realizando el usuario en función de la Tabla 10.

Figura 63 Etiquetado de un usuario andando

Figura 64 Etiquetado de un usuario corriendo

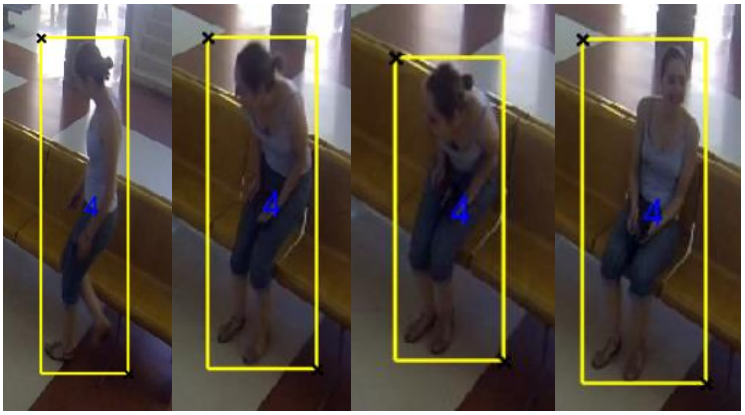


Figura 65 Etiquetado de un usuario sentándose

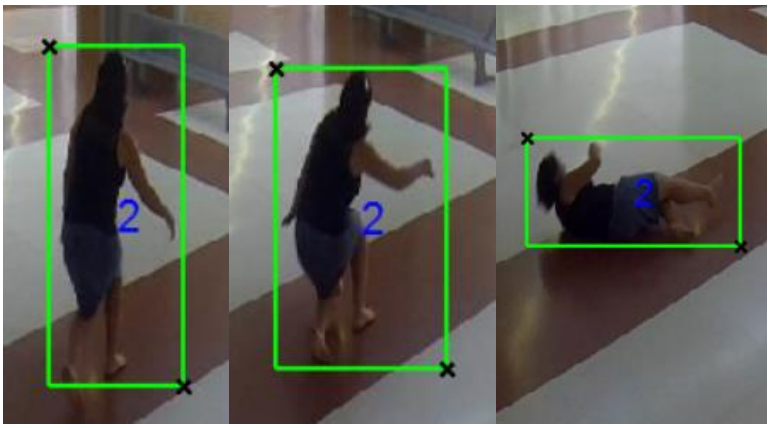


Figura 66 Etiquetado de una persona cayéndose

Figura 67 Ejemplo de un usuario sin identificador sin realizar acción alguna

En la Tabla 11 se muestran el número de secuencias para cada acción además del número de actores realizándolas.

Tabla 11 Personas y secuencias realizando distintas acciones

ACCIÓN	PERSONAS	SECUENCIAS
Andar	17	52
Correr	12	24
Sentarse	9	9
Caerse	9	17

Las secuencias pueden ser de dos tipos, simples o complejas. Las secuencias simples incluyen solo a un usuario realizando una acción concreta, como por ejemplo el usuario con identificador 4 cayéndose al suelo (ver Figura 68).

Las secuencias complejas en cambio, son aquellas que engloban distintos usuarios realizando diferentes acciones de manera simultánea, como por ejemplo varios usuarios con distintos identificadores sentándose mientras otros van andando (ver Figura 69).

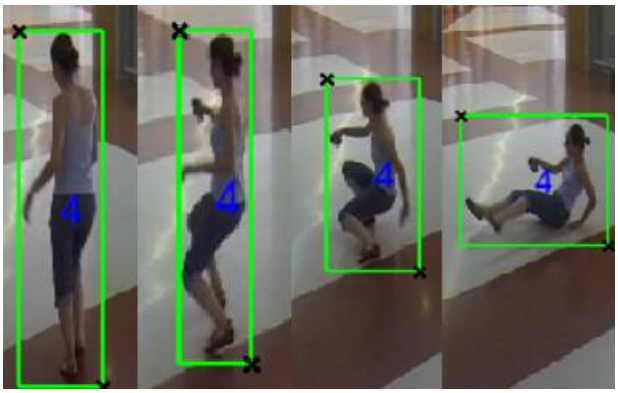


Figura 68 Ejemplo de secuencia simple

Figura 69 Ejemplo de secuencia compleja

En la Tabla 12 se pueden ver las secuencias complejas y simples en función de las acciones realizadas, siendo noventa y seis el total de todas las secuencias.

Tabla 12 Resumen de secuencias complejas y simples

ACCIÓN	Secuencias simples	Secuencias complejas
Andar	41	5
Correr	23	2
Sentarse	5	4
Caerse	15	1

En la Tabla 13 se puede ver el resultado final de la base de datos obtenida, los vídeos grabados, las secuencias en cada uno de ellos con sus usuarios correspondientes, y toda la información necesaria.

Tabla 13 Tabla resumen del resultado de la base de datos

VÍDEO	NOMBRE	DURACIÓN (segundos)	FRAMES	USUARIOS	ACCIÓN	SECUENCIA
Vídeo 1	GOPR0472.MP4	0:03:57	11858	5	Andar	5 simples
Vídeo 2	GOPR0473.MP4	0:01:15	3756	4	Correr	4 simples
Vídeo 3	GOPR0474.MP4	0:00:46	2300	2	Correr	2 simples
Vídeo 4	GOPR0475.MP4	0:03:26	10332	5	Andar y sentarse	5 simples
Vídeo 5	GOPR0476.MP4	0:01:15	3755	4	Andar, sentarse y correr	2 complejas
Vídeo 6	GOPR0477.MP4	0:00:02	139	-	Vídeo de fondo	-
Vídeo 7	GOPR0478.MP4	0:00:45	2275	2	Andar y caerse	2 simples
Vídeo 8	GOPR0479.MP4	0:00:24	1208	1	Andar y caerse	1 simples
Vídeo 9	GOPR0009.MP4	0:04:24	13200	5	Andar	6 simples
Vídeo 10	GOPR0013.MP4	0:02:42	8100	6	Andar	6 simples
Vídeo 11	GOPR0014.MP4	0:01:29	4450	5	Correr	5 simples
Vídeo 12	GOPR0015.MP4	0:00:40	2000	4	Correr	4 simples
Video 13	GOPR0016.MP4	0:02:50	8500	4	Andar	5 simples
Video 14	GOPR0017.MP4	0:01:12	3600	4	Correr	4 simples
Video 15	GOPR0018.MP4	0:01:40	5000	4	Andar	4 simples
Video 16	GOPR0019.MP4	0:01:08	3400	4	Correr	4 simples
Video17	GOPR0169.MP4	0:01:34	1676	5	Andar y caerse	6 simples
Video18	GOPR0170.MP4	0:01:05	1170	5	Andar y caerse	5 simples
Video19	GOPR0171.MP4	0:02:45	9949	6	Andar, sentarse y caerse	3 complejas y 2 simples

Capítulo 5

Conclusiones y trabajos futuros

En los capítulos anteriores se ha realizado una descripción del planteamiento del problema, las soluciones adoptadas y los resultados obtenidos.

En este capítulo se presentarán las conclusiones alcanzadas tras la realización del trabajo, y se expondrán las posibles mejoras y futuras líneas de trabajo.

5.1. Conclusiones

Tras realizar este TFG se desarrollan las siguientes conclusiones:

- El reconocimiento de la actividad humana es, y seguirá siendo, un campo de la inteligencia artificial relevante y bastante estudiado debido a la infinidad de aplicaciones que puede tener. El número de publicaciones al respecto es muy elevado y día a día sigue aumentando.

- Con respecto a la metodología de grabación, la tendencia es utilizar cámaras RGB debido a la gran cantidad de información que proporciona, y su facilidad a la hora de adaptarlas a diferentes aplicaciones. Es cierto que poco a poco se está empezando a trabajar con cámaras RGBD (color y profundidad), además de los sistemas basados en mapas de profundidad a través de la Kinect, pero siempre y cuando no aumente la complejidad de la grabación y adquisición de datos.

- Existen infinidad de acciones estudiadas, casi todas ellas complejas, pero se ha comprobado que cada base de datos es específica para cada aplicación, lo que hace necesario la creación de nuevas bases de datos más realistas y menos concretas.

- Los escenarios de grabación normalmente son homogéneos, siendo muchas veces un fondo estático sin posibilidad de que aparezcan variables dinámicas, lo que hace que exista la necesidad de ampliar las bases de datos e incluir la posibilidad de que aparezcan situaciones inesperadas.

- En este proyecto se ha mejorado la interfaz de usuario existente para la fase de anotación, siendo ahora más intuitiva, con nuevas opciones para aportar más información al usuario. Además, se ha incluido todo lo relacionado con las acciones realizadas por el usuario, la distinción entre las acciones a través de los colores y su inclusión dentro de los archivos GT

- La inclusión de la acción en dicho reconocimiento y su diferenciación mediante colores permite obtener más datos acerca de los usuarios, siendo

importante sobre todo para aplicaciones de video-vigilancia, donde la seguridad es un pilar fundamental.

- Se ha conseguido crear una base de datos completa con nuevas actividades estudiadas y bien documentada, a partir de la cual se puede proceder a realizar las pruebas oportunas para la validación de los algoritmos de reconocimiento.

- Se ha podido deducir la necesidad de ampliar el número de secuencias de cara a realizar un mejor entrenamiento, y ampliar el número de actividades a estudiar.

5.2. Trabajos futuros

A partir del desarrollo de este TFG se proponen las siguientes mejoras o modificaciones futuras:

- Realizar una ampliación del número de secuencias de vídeo para un mejor entrenamiento de la máquina de detección, incluyendo nuevos escenarios de grabación, incluso utilizar los mismos escenarios, pero con puntos de vista diferentes.

- Utilizar nuevos usuarios con distintas características, como por ejemplo más mujeres, distintas edades, distintas estaturas, etc. Incluso se podría incluir grabaciones con niños debido a que tanto la fisionomía como la manera de andar o correr es totalmente diferente a la de un adulto.

- Incluir clases de acciones nuevas a reconocer, como por ejemplo la interacción entre personas o acciones diferentes, lo que implicaría aumentar el número de secuencias a grabar

- Emplear otras técnicas de grabación, como por ejemplo aplicar la técnica RGBD para incluir datos de profundidad y mejorar así el funcionamiento del sistema, o añadir nuevos datos

- Mejorar la interfaz de usuario incluyendo nuevas características a reconocer, como por ejemplo posiciones relativas a distintos puntos de interés

- Realizar un ejecutable para la anotación y así no tener que ejecutarlo sobre MATLAB, de esta manera puede ser exportable a cualquier ordenador y ser utilizada por cualquier usuario sin conocimientos sobre MATLAB.

- Aparte de la generación de archivos GT donde se almacena la información de la anotación, generar vídeos en formato .mp4 con el resultado del etiquetado

Apéndice A

Presupuesto

En esta parte del proyecto se hace una estimación del coste total que supondría la ejecución del mismo, utilizando los mismos recursos que en este caso. En los apartados siguientes aparecen los gastos agrupados según su origen, y en el último apartado se detalla el presupuesto total.

1. Recursos hardware

Los recursos hardware que se han utilizado en este proyecto son los resumidos en la Tabla 14

Tabla 14 Resumen económico de los recursos HW usados

CONCEPTO	PRECIO UNIT.	CANTIDAD	SUBTOTAL
Ordenador portátil ASUS GL552JX, i7-4720Q	1149,00€	1	1149,00€
Cámara GoPro HERO	139,99€	1	139,99€
Trípode tipo pulpo	8,99€	1	8,99€
		SUBTOTAL	**1297,98€**

2. Recursos software

Los recursos software utilizados en este proyecto son los resumidos en la Tabla 15

Tabla 15 Resumen económico de los recursos SW usados

CONCEPTO	PRECIO UNIT.	CANTIDAD	SUBTOTAL
Microsoft Office Word 2013	60,00€	1	60,00€
Windows 8.1	180€	1	180€
Licencia MATLAB “Education”	500€	1	500€
		SUBTOTAL	**740€**

3. Coste de la mano de obra

Tabla 16 Resumen económico del coste de la mano de obra

CONCEPTO	PRECIO UNIT.	CANTIDAD	SUBTOTAL
Desarrollo SW	65€/h	250	16.250,00€
Redacción de la memoria	15€/h	50	750€
		SUBTOTAL	**17.000,00€**

4. Presupuesto de ejecución material

Es la suma total de los importes del coste de materiales y de la mano de obra.

Tabla 17 Coste total del proyecto

CONCEPTO	PRECIO
Coste recursos hardware	1297,98€
Coste recursos software	740,00€
Coste mano de obra	17.000,00€
TOTAL	**19.037,98€**

El presupuesto total del proyecto asciende a la cantidad de diecinueve mil treinta y siete con noventa y ocho céntimos.

Alcalá de Henares a 16 de septiembre de 2016.

Firmado: Valeria Boggian Arévalo

Graduada en Ingeniería Electrónica y Automática Industrial

Apéndice B

Manual de usuario

1. Requisitos de la aplicación

Para la utilización del software de etiquetado es necesario tener instalado Matlab 2013b o versiones inferiores.

2. Interfaz gráfico

Para iniciar la interfaz de usuario, en primer lugar es necesario arrancar el programa MATLAB, y acceder a la carpeta donde se encuentran los ficheros .m y .fig. Otra opción es ejecutar el archivo *annotation_interface.m*, o bien escribir en la línea de comandos:

```
>> annotation_interface
```

Esta interfaz de usuario está compuesta por dos figuras, una figura de anotación (Figura 70), donde se realiza el etiquetado de los vídeos, y una figura de comandos (Figura 71), donde se puede encontrar una serie de pantallas informativas, así como todos los botones necesarios para realizar la anotación de la manera más sencilla posible.

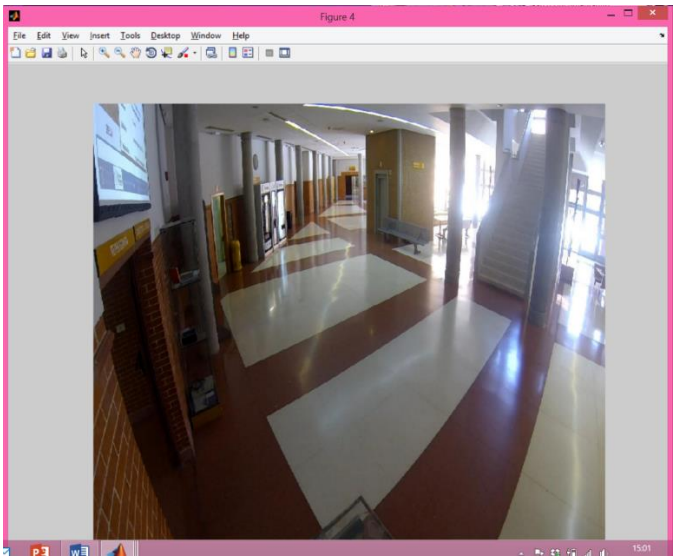


Figura 70 Ventana de anotación

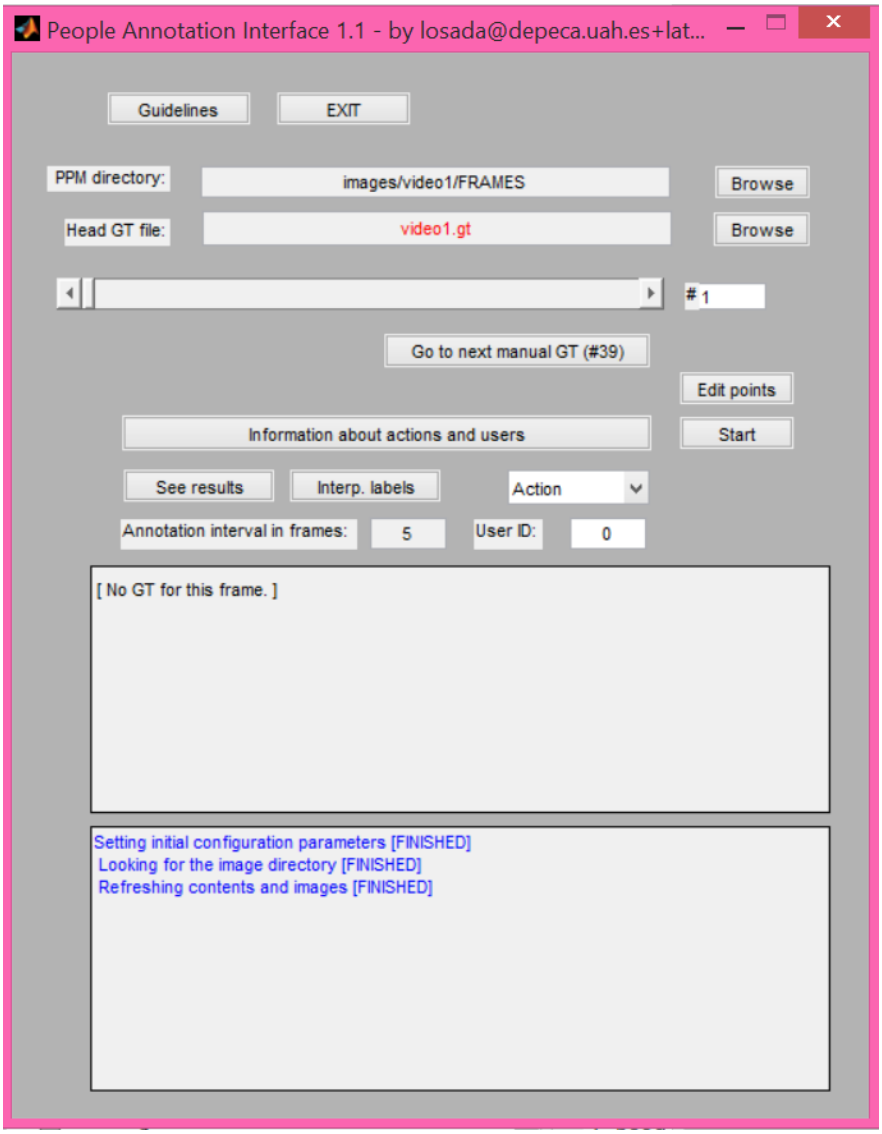


Figura 71 Ventana de comandos

Dentro de la ventana de comandos encontramos las siguientes características:

- Comandos de elección de directorio
  Estos comandos, distribuidos en la parte superior de la ventana, se componen de dos botones de navegación por los distintos directorios, y de dos barras con el nombre del directorio y archivo GT elegido. Estos comandos sirven para elegir el vídeo que se quiere anotar.

Figura 72 Comandos de elección de directorio

- Botones de navegación

  Este conjunto de botones está formado por un *slider* que sirve para navegar a lo largo de todos los *frames* del vídeo, así como para ver en qué punto del vídeo nos encontramos. Al lado de este *slider* se puede ver el número del *frame* que se está mostrando en la pantalla de anotación. Además, debajo del *slider* hay dos botones para ir al siguiente *frame*, o al previo, etiquetado manualmente.

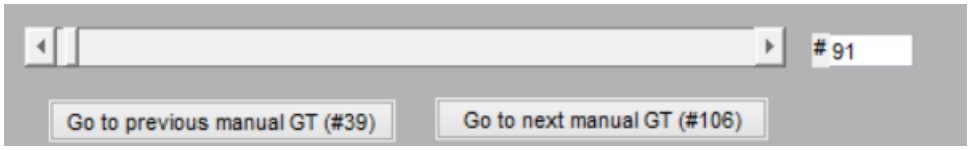


Figura 73 Botones de navegación

- Botones de control de anotación

  Son aquellos que sirven para realizar el proceso de etiquetado:

  - *START*

    Botón que sirve para comenzar el proceso de anotación

Figura 74 Botón START

  - *EDIT POINTS*

    Botón para editar los puntos que se han etiquetado previamente

Figura 75 Botón EDIT POINTS

  - *INTERP. LABELS*

    Botón que sirve para completar el etiquetado manual y que se anoten aquellos *frames* que no han sido etiquetados por el usuario mediante interpolación lineal.

Figura 76 Botón INTERP. LABELS

- *SEE RESULTS*
  Botón que sirve para ver en la imagen de anotación los resultados del etiquetado de un vídeo

Figura 77 Botón SEE RESULTS

- Comandos para el establecimiento de parámetros
  Este conjunto de opciones sirve para que el usuario pueda elegir el identificador del actor que aparece en el vídeo, la acción que realiza y el intervalo que desea que haya entre un *frame* y otro en el proceso de anotación manual. Cuanto menor sea este intervalo, mayor precisión tendrán los resultados.

Figura 78 Establecimiento de parámetros

En este proyecto las acciones que han sido elegidas para ser etiquetadas son andar, correr, caerse, sentarse y quedarse quieto, por ello estas son las opciones que nos proporciona el desplegable de la acción.

Figura 79 Desplegable de la acción

- Opciones de información
  Para que esta interfaz sea de ayuda e intuitiva, se han diseñado una serie de elementos que sirven para que el usuario sepa cómo actuar. En primer lugar, en la parte superior de la ventana aparece un botón, *GUIDELINES*, que, al pulsarlo, te lanza un cuadro de diálogo donde explica brevemente la función de la aplicación.

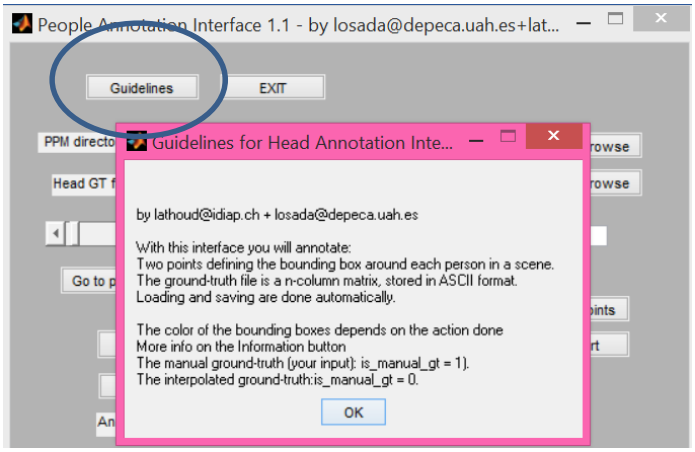


Figura 80 Botón GUIDELINES

En segundo lugar, en la zona de los comandos para el establecimiento de parámetros, hay un botón que sirve para obtener información sobre los usuarios, es decir, qué identificador tiene cada persona en el vídeo, y sobre acciones, descifrando el color al que está asociado cada acción.

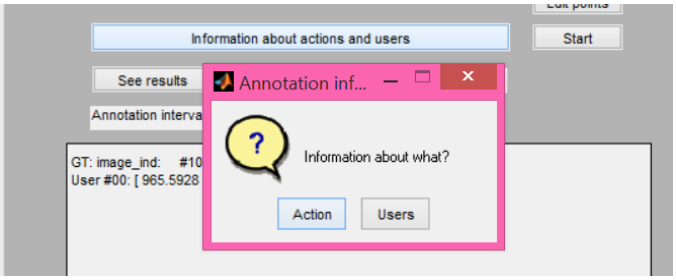


Figura 81 Botón de información

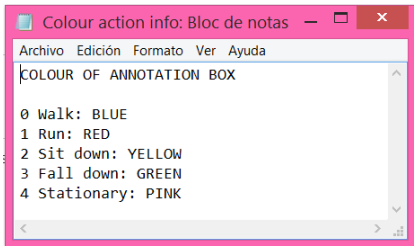


Figura 82 Información sobre los colores de la anotación

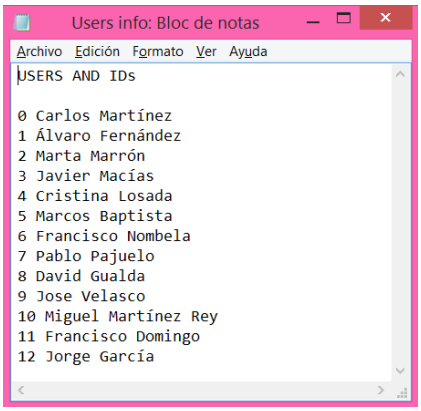


Figura 83 Información sobre los identificadores de usuario

Por último, en la parte inferior de la interfaz, existen dos ventanas blancas donde aparece información relativa al proceso de anotación. En la ventana superior se proporciona información sobre los archivos GT, es decir, informa al usuario de si el *frame* seleccionado ha sido etiquetado (Figura 85), en cuyo caso presenta el archivo GT, o si no lo ha sido (Figura 84).

Figura 84 *Frame* no etiquetado

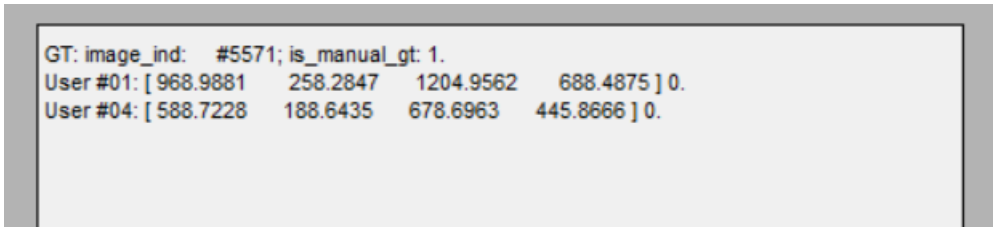


Figura 85 *Frame* etiquetado con dos usuarios

En la parte inferior aparece otra ventana blanca donde se puede ver un breve resumen del proceso cuando se abre el programa (Figura 86), o bien unas breves indicaciones sobre qué teclas se deben utilizar para realizar el proceso de anotación (Figura 87)

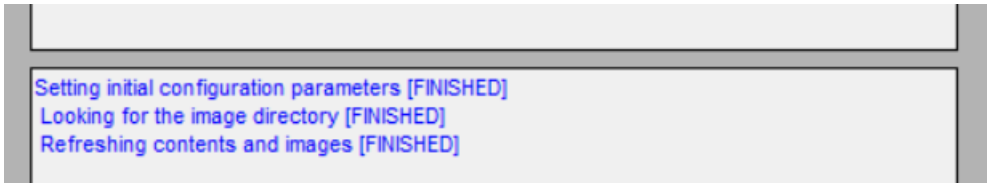


Figura 86 Proceso de apertura de la interfaz

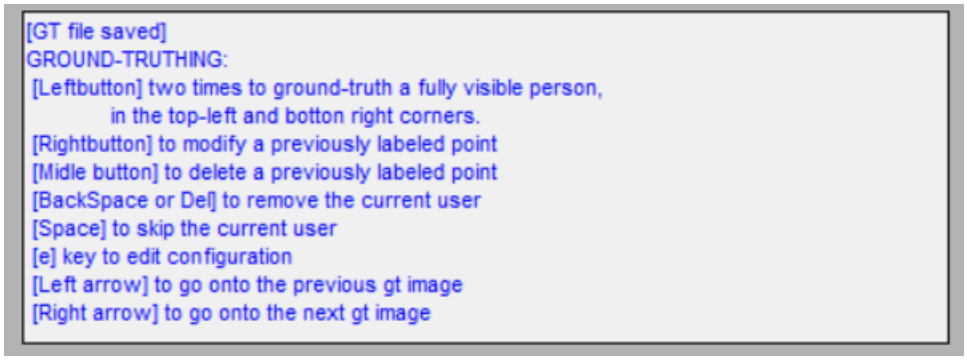


Figura 87 Información sobre cómo se debe realizar la anotación

- Botón *EXIT*
  Botón para salir del programa de anotación

2.1. Selección del directorio de imágenes y fichero de ground truth

En la parte superior de la figura *annotation_interface.fig,* aparecen dos campos editables y dos botones que permiten seleccionar la carpeta donde se encuentran los archivos de datos (*PPM directory*) y el fichero donde se almacenará el *ground truth* (GT) (*Head GT file*) tras el etiquetado.

Al iniciar el programa se comprueba si en la carpeta de trabajo existe una carpeta denominada *images* que contenga algún directorio.

En caso afirmativo, se comprueba si contiene imágenes .jpg y se elige como directorio. Además, se asigna al fichero de GT el mismo nombre del directorio y la extensión .gt.

Si no se encuentra ningún directorio, o bien si se desea cambiar de carpeta, existen varias alternativas:

- Cambiar de carpeta escribiendo el *path* y el nombre del nuevo directorio en el campo editable.
- Elegir una nueva carpeta pulsando el botón *BROWSE* y seleccionando un fichero perteneciente al nuevo directorio.

De la misma manera, es posible cambiar el fichero de GT tanto escribiendo el nombre en el campo correspondiente, como pulsando el botón *BROWSE,* y seleccionando el nuevo archivo.

Dentro del directorio de imágenes pueden encontrarse tanto los ficheros de datos de cada una de las imágenes de una secuencia dada, como el video (.avi o .mp4) que contiene la secuencia completa.

En caso de que no existan los ficheros asociados a cada imagen individual, el programa los extraerá y almacenará en el mismo directorio, dentro de una carpeta denominada FRAMES. Si el número de imágenes es elevado, este proceso puede tardar varios minutos.

Una vez elegido el directorio, se mostrará la primera imagen de la secuencia.

Desplazando el *slider* que se muestra en la Figura 88, es posible representar los diferentes ficheros de datos, y el etiquetado en caso de estar disponible.

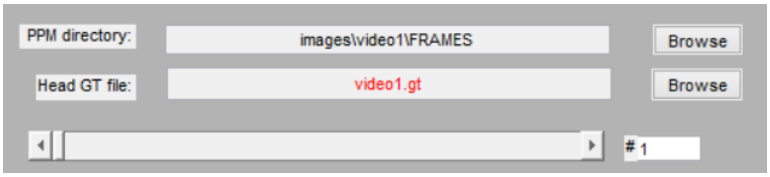


Figura 88 Visualización de la carpeta de ficheros y archivo de datos

2.2. Configuración

A continuación, en la parte central de la interfaz se puede editar el intervalo entre anotaciones, es decir, cada cuántas imágenes se realiza la anotación. Para ello se puede editar el campo *Annotation interval in frames* (ver Figura 89).

También se puede editar manualmente el identificador de usuario (*UserID*) y elegir la acción realizada de un desplegable con las distintas opciones (*Action*)

Figura 89 Botones de configuración

Finalmente, en esta sección es posible iniciar diferentes procesos:

- El etiquetado de los puntos pulsando el botón *Start*.
- La edición de puntos editados previamente, mediante el botón *Edit Points*.
- La visualización del etiquetado, utilizando el botón *See Results*.

3. Proceso de etiquetado (*START BUTTON*)

Una vez configurados los diferentes parámetros, puede iniciarse el proceso de etiquetado pulsando el botón *Start*. En ese momento, la información del panel inferior cambiará, mostrando las instrucciones para el etiquetado, tal como se muestra en la Figura 90, y al pasar el cursor sobre la imagen este cambiará convirtiéndose en dos ejes.

Si el usuario ha sido etiquetado previamente en esa imagen, aparecerá destacado en color blanco.

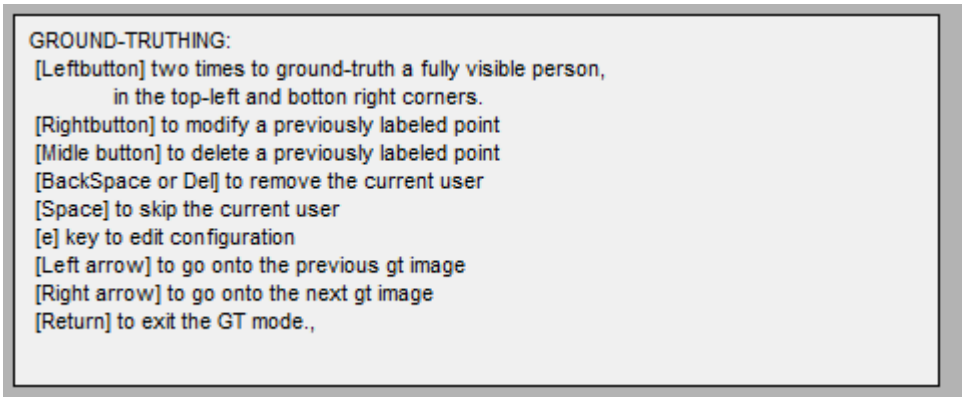


Figura 90 Instrucciones de etiquetado en el panel de información

3.1. Etiquetado del usuario

En la situación anterior es posible etiquetar al usuario configurado en el campo *UserID*.

- Si se hace *click* con el botón izquierdo del ratón sobre la imagen dos veces (una en la esquina superior izquierda y otra en la inferior derecha), se etiquetará al usuario.
- Si el usuario ya había sido etiquetado, se sustituyen los datos anteriores.

- Si algún punto se ubica en un lugar no deseado, es posible eliminarlo haciendo *click* sobre él con el botón central del ratón. Sólo podrán eliminarse los puntos pertenecientes al usuario seleccionado.

3.2. Edición del etiquetado previo

Existe la opción de modificar los puntos que han sido etiquetados previamente para el usuario seleccionado.

- Si se hace *click* con el botón derecho del ratón sobre la imagen, se inicia el modo de edición. En este caso cambiará el panel de información mostrando las instrucciones para editar los puntos (Figura 91)
- Haciendo *click* con el botón izquierdo del ratón se modificará la ubicación del punto más próximo perteneciente al usuario seleccionado.
- La edición se finaliza haciendo de nuevo *click* con el botón derecho del ratón sobre la imagen.

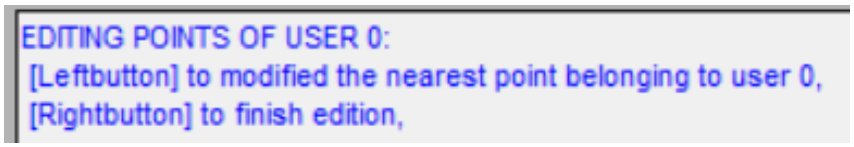


Figura 91 Instrucciones para la edición de puntos durante el proceso de etiquetado

3.3. Eliminación de un usuario

Un usuario etiquetado puede eliminarse pulsando la tecla *supr* o *delete*.

3.4. Avance o retroceso en la secuencia de imágenes

Las teclas *flecha izquierda* y *flecha derecha* del teclado permitirán respectivamente avanzar o retroceder el número de imágenes configurado en el campo *Annotation interval in frames* dentro de la secuencia.

3.5. Modificación de la configuración durante el etiquetado

Durante el etiquetado es posible modificar tanto el intervalo entre *frames* etiquetados como el identificador de usuario o la acción a realizar. Para ello se debe pulsar la tecla e y editar el campo correspondiente en menos de 30 segundos.

3.6. Finalización del etiquetado

La tecla *Enter* finaliza el proceso e etiquetado.

4. Proceso de edición (EDIT POINTS BUTTON)

Pulsando el botón *Edit points* es posible modificar los puntos etiquetados anteriormente. Cuando se pulsa ese botón aparece una serie de instrucciones a seguir.

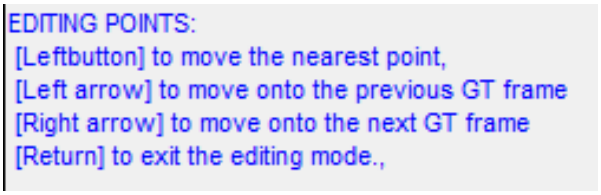


Figura 92 Información para el usuario de cómo realizar la edición de puntos

Esta funcionalidad es similar a la explicada anteriormente, sin embargo, permite la edición de los puntos pertenecientes a cualquiera de los usuarios, no solo al configurado en *UserID*.

En este modo, se tienen las siguientes opciones:

- Haciendo *click* con el botón izquierdo del ratón se modifica la posición del punto que se encuentre más próximo.
- Las teclas *flecha izquierda* y *flecha* derecha permiten recorrer la secuencia de imágenes de una en una.
- La tecla *Enter* finaliza el proceso de edición.

5. Fichero GROUND TRUTH

Tras el etiquetado de cada secuencia de imágenes se genera un fichero de *ground-truth* con la información resultante.

Por defecto, este fichero tendrá el mismo nombre que la secuencia de imágenes y la extensión .gt (aunque puede ser modificado en la interfaz, tal como se ha descrito en el apartado 2.1).

En el fichero generado se almacena una línea por cada imagen etiquetada, que contiene una serie de valores numéricos separados por espacios, con el formato definido en el fichero20140722-PeopleCountingDatabase-Design-v2.docx:

$$FFFFFF\ X\ UID_0\ PAX_0\ PAY_0\ PBX_1\ PBY_1\ ACT_0 \ldots\ UID_i\ PAX_i\ PAY_i\ PBX_i\ PBY_i\ ACT_i$$

Donde:

- FFFFFF: es el número de imagen dentro de la secuencia. Este valor es especialmente importante en los casos en que no se etiquetan todas las imágenes, sino algunas de ellas.
- X: indica si el etiquetado de esa imagen es manual (1) o automático (0) (mediante interpolación entre las etiquetas manuales).
- Número de usuario, coordenadas de los puntos etiquetados y acción realizada por el usuario: a continuación, por cada usuario etiquetado en la imagen, aparecen 6 valores. El primero identifica al usuario (UID_0), los siguientes 4 corresponden a las coordenadas (en píxeles) de los puntos etiquetados sobre la imagen y el último valor corresponde al identificador de la acción realizada (ACT_0).

En la Figura 93 se muestra un ejemplo de las líneas obtenidas para las imágenes entre la imagen 91 y la 106 de una secuencia, en la que ha etiquetado al usuario 0 manualmente en la primera y última imagen, y de forma automática en las intermedias.

```
000091 1 0000  1053.8687   274.0101  1247.3966   678.3783 0
000092 0 0000  1047.9836   273.4859  1242.0774   678.6778 0
000093 0 0000  1042.0986   272.9617  1236.7582   678.9774 0
000094 0 0000  1036.2135   272.4376  1231.4390   679.2769 0
000095 0 0000  1030.3285   271.9134  1226.1199   679.5764 0
000096 0 0000  1024.4434   271.3892  1220.8007   679.8760 0
000097 0 0000  1018.5583   270.8650  1215.4815   680.1755 0
000098 0 0000  1012.6733   270.3408  1210.1623   680.4750 0
000099 0 0000  1006.7882   269.8167  1204.8431   680.7746 0
000100 0 0000  1000.9032   269.2925  1199.5239   681.0741 0
000101 0 0000   995.0181   268.7683  1194.2047   681.3736 0
000102 0 0000   989.1330   268.2441  1188.8855   681.6732 0
000103 0 0000   983.2480   267.7199  1183.5664   681.9727 0
000104 0 0000   977.3629   267.1958  1178.2472   682.2722 0
000105 0 0000   971.4779   266.6716  1172.9280   682.5718 0
000106 1 0000   965.5928   266.1474  1167.6088   682.8713 0
```

Figura 93 . Ejemplo de líneas almacenadas en el fichero de ground-truth

Apéndice C

Pliego de condiciones

Para la correcta utilización del sistema desarrollado en este trabajo se debe disponer de un hardware y un software que cumpla unos requisitos mínimos.

1. Requisitos de hardware

- Procesador de 32 o 64 bits
- 2G de memoria RAM o superior
- Al menos 300MB de memoria libres en el disco duro para funciones y datos
- Al menos 16GB de memoria libres en el disco duro para los vídeos de la base de datos y las funciones de MATLAB
- Cámara de vídeo GoPro HERO4 o similar.

2. Requisitos de software

- Sistema operativo Linux o Windows 7 (o superiores)
- Al menos 300MB de memoria libres en el disco duro para funciones y datos
- Tener instalado MATLAB 2013b o versiones inferiores

Referencias

[1] Miguel A. Realpe, Boris X. Vintimilla, Dennis G. Romero, Paolo Remagnino. "Análisis de comportamiento humano: Metodología para localización y seguimiento de personas en secuencias de video." Facultad de Ingeniería en Electricidad y Computación, Universidad de Guayaquil, Ecuador, 2010.

[2] Página de la base de datos de "*Idiap Head Pose Database*," https://www.idiap.ch/dataset/headpose [Último acceso 13/septiembre/2016].

[3] Carlos Martínez García. "Reconocimiento de actividad humana en secuencias de vídeo". Trabajo fin Máster en la Universidad de Alcalá, 2015.

[4] M. Edwards, J. Deng, X. Xie. "From Pose to Activity: Surveying Datasets and Introducing CONVERSE", Universidad de Swansea, 2015.

[5] Pedro Nogueira. "Motion Capture Fundamentals: A critical and comparative analysis on real-world applications", Facultad de Ingeniería de la Facultad de Oporto, 2011

[6] M. Andersen, T. Jensen, P. Lisouski, A. Mortensen, M. Hansen, T. Gregersen, P. Ahrendt. "Kinect depth sensor evaluation for computer vision applications," Universidad de Aarhus, 2012.

[7] L. Li, T. Nawaz, J. Ferryman. "PETS 2015: Datasets and Challenge, Proceedings of the IEEE International Conference on Advanced Video- and Signal-based Surveillance (AVSS)", Karlsruhe, 2015.

[8] S. Blunsden, R. B. Fisher. "THE BEHAVE VIDEO DATASET, Annals of the BMVA", Universidad de Edinburgo, 2010.

[9] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal and R. Bajcsy. "Berkeley MHAD: A Comprehensive Multimodal Human Action Database." In Proceedings of the IEEE Workshop on Applications on Computer Vision (WACV), 2013. http://tele-immersion.citris-uc.org/berkeley_mhad [Último acceso 13/septiembre/2016]

[10] La página web oficial de CAVIAR: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/ [Último acceso 13/septiembre/2016]

[11] Leonid Sigal and Michael J. Black. "HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion", 2006

[12] D. Weinland, M. Ozuysal, P. Fua. “Making Action Recognition Robust to Occlusions and Viewpoint Changes”, Laboratorios Telekom, 2006

[13] D. Weinland, R. Ronfard, E. Boyer. “Free Viewpoint Action Recognition using Motion History Volumes”, INRIA, 2006

[14] Página web de KTH: http://www.nada.kth.se/cvap/actions/ [Último acceso 13/septiembre/2016]

[15] C. Schüldt, I. Laptev, B. Caputo. “Recognizing Human Actions: A Local SVM Approach”, Real Instituto de Tecnología de Estocolmo, 2014

[16] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri. “Actions as Space-Time Shapes”. Instituto de ciencias de Weizmann, 2007

[17] Sileye O. Ba, J. M. Odobez. “A video database for head pose tracking evaluation”. IDIAP Research Institute, 2005.

[18] D. Orlando Barragán. “Manual de interfaz gráfica de usuario en MATLAB”. 2008

[19] Página oficial de GoPro: https://es.gopro.com/ [Último acceso 13/septiembre/2016]

[20] C. Martínez, M. Baptista, C. Losada, M. Marrón, V. Boggian. “Human action recognition in realistic scenes base on Action Bank”, Grupo GEINTRA, Universidad de Alcalá. 2016

[21] L. Li,T. Nawar, J. Ferryman. “PETS 2015: Datasets and challenges”, grupo de Vision Computacional, Universidad de Reading, UK, 2015

[22] Página web: http://www.rubberonion.com/is-performance-capture-live-action-or-animation/ [Último acceso 15/septiembre/2016]

Universidad de Alcalá
Escuela Politécnica Superior

ESCUELA POLITECNICA
SUPERIOR

Universidad
de Alcalá