Capítulo 4
Interfaz de Control Domótico Basada en un Sistema de Detección de Postura

Francisco Flórez Revuelta y Alexandros Andre Chaaraoui

Resumen Este artículo describe la especificación y el diseño de una interfaz de control domótico basada en un sistema de detección de postura por visión. Mediante las imágenes captadas por una cámara cenital se detecta la dirección y el sentido de la postura del usuario empleando varias técnicas de inteligencia artificial basadas en análisis de componentes principales, clasificadores débiles y redes neuronales complejas. Esta información se obtiene para ser combinada en trabajos posteriores con comandos de voz con el objetivo de permitir una interacción persona-entorno inteligente.

4.1. Introducción

En el campo del hogar digital disponemos de tecnologías que son capaces de reconocer correctamente determinados comandos de voz y realizar acciones previamente configuradas en consecuencia. De este modo, el usuario puede interactuar con los dispositivos de su vivienda a través de su voz. Comandos como “bajar la persiana del salón” o “abrir la puerta del baño” realizan la acción esperada dadas las características de la vivienda domotizada. ¿Y si el usuario pudiera decir solo “bajar la persiana” o “ábrete” y el sistema fuera capaz de reconocer a qué objeto hace referencia? Es aquí, donde entran en juego las técnicas de visión necesarias para reconocer la postura de la persona, de forma que se pueda determinar hacia dónde está mirando y establecer de este modo qué persiana se debe bajar, o qué puerta se debe abrir. Éste es el problema que se pretende resolver en este trabajo.

Francisco Flórez Revuelta
Departamento de Tecnología Informática y Computación, Universidad de Alicante
e-mail: florez@dtic.ua.es

Alexandros Andre Chaaraoui
Departamento de Tecnología Informática y Computación, Universidad de Alicante
e-mail: alexandros@dtic.ua.es

Es lógico pensar que en un entorno inteligente el hecho de tener que indicar que queremos encender la luz número 3 del salón no es la opción más amigable, ni tampoco es una forma muy intuitiva. El usuario desea encender una luz concreta, y es muy probable que en ese momento esté dirigiendo su mirada hacia ella.

La investigación en interfaces humano-máquina (HCI) tiene como objetivo desarrollar algoritmos y sistemas que detecten y perciban a las personas y su actividad, permitiendo elaborar interfaces de computador más naturales, intuitivas y efectivas. Intuitivamente, vemos que los aspectos visuales que empleamos en la comunicación entre las personas, como el reconocimiento de la identidad, la edad, la postura, la dirección de la mirada y nuestros gestos, deberían ser los elementos a emplear en la interacción humano-máquina, ya que éstos constituyen la forma natural con la que nos comunicamos. Es por esto que, aunque también se consideran interfaces por voz y por tacto, las más estudiadas son las que emplean la visión. Así llegamos al área de investigación en la interacción humano-máquina basada en visión (*Vision based Human-Computer Interaction*). Este área es parte de un contexto más amplio que conocemos como interfaces perceptuales, interfaces multimodales e interfaces post-WIMP [1], las cuales intentan integrar múltiples modalidades perceptuales como la visión, el habla, los gestos y la háptica. En definitiva, el objetivo de estas áreas de investigación es superar la era de las interfaces gráficas de usuarios (GUI) y los periféricos tradicionales, con el objetivo de adecuarse a los entornos de computación futuros, como son, entre otros, los entornos domóticos.

4.2. Estado de la cuestión y técnicas de resolución disponible

4.2.1. Análisis del estado de la cuestión

Al revisar las técnicas de resolución disponibles actualmente, vemos que se trata de un campo que se encuentra, en gran parte, sin resolver, ya que, como veremos, en los métodos propuestos abundan restricciones técnicas y limitaciones del ámbito de uso de los resultados.

Existen técnicas de visión capaces de deducir hacia dónde mira una persona basándose en obtener primero la ubicación de sus ojos y de las pupilas, para posteriormente obtener la dirección de la mirada a partir del plano de la cara con relación a la posición de las pupilas [2]. En otros casos, se logra obtener la dirección hacia la que está enfocado el cuerpo de la persona partiendo de una imagen completa de su cuerpo y calculando la posición relativa de los brazos y/o de los pies [3] [4]. En [5] se emplea el movimiento del blob detectado para calcular la dirección en la que se mueve el individuo y se asume que ésta también es la dirección en la que miraba. Por lo contrario, en [6] se parte de un reconocimiento facial y se calcula la posición relativa de ésta con respecto al cuerpo para obtener su orientación.

Existen diversas aplicaciones que requieren de una detección de postura con mayor o menor precisión y que están basadas en la orientación del cuerpo del individuo,

de la cabeza, o incluso de la propia mirada. Así encontramos, por ejemplo, en [7] que se han realizado estudios sobre los factores determinantes entre la relación de estos tres tipos de posturas y sus contribuciones a la postura global. En este caso se trata de realizar un análisis sobre el comportamiento de las personas en reuniones de más de dos individuos. La ventaja de emplear una detección de postura basada en la orientación del cuerpo o de la cabeza frente a la detección de los ojos consiste principalmente en que la primera se puede realizar de una forma menos intrusiva y, por tanto, también es apta para entornos públicos.

En [3], vemos que a partir del trabajo anterior, se pasó a integrar la detección y el *tracking* de manos y brazos junto al de la cabeza para obtener mejores resultados en la detección de la postura global y abrir el camino a nuevas interfaces para la interacción humano-robot. En concreto, el objetivo era poder indicarle al robot qué debe hacer dando la indicación con el habla y con el gesto de señalar un lugar, como si de una mascota se tratase. Se trata por tanto de una aplicación similar a la de este proyecto, sobre todo porque se parte de la misma premisa, que consiste en que las personas tienden a dirigir la mirada hacia los objetos con los que interactúan, hecho que avalan varios estudios [8] [9]. Cabe destacar que este sistema requiere de una cámara estéreo para localizar las partes del cuerpo humano y poder iniciar el *tracking*. En concreto, se emplearon modelos ocultos de Markov para reconocer los gestos de las manos y combinarlos con la detección de la orientación de la cabeza, obteniendo así una conclusión final sobre la dirección de la postura de la persona. En cuanto a la detección de la dirección de la postura, se parte de histogramas de color para poder detectar la piel en la imagen 3D y reconocer las manos y la cabeza. A continuación, se combina la dirección resultante de alinear cabeza y manos con la dirección del brazo y la de la cabeza. Mientras que las dos primeras direcciones se obtienen a partir de la imagen 3D, la última emplea un sensor magnético. Mediante este sistema se logra obtener una tasa de acierto de entre el 74 y el 83,1 %.

En trabajos más recientes, como es el caso de [10], podemos ver que la investigación mencionada ahora también se aplica en entornos domóticos y se ha llegado a mejorar su tasa de acierto hasta alcanzar un 87,8 %. Sin embargo, en este entorno el sistema ha sufrido varios cambios al basarse únicamente en visión 2D, pero con imágenes procedentes de múltiples cámaras. El método de resolución utilizado consiste en reconocer la dirección de la parte superior del cuerpo empleando solamente su silueta. Este dato se clasifica mediante histogramas de contexto de forma [1] con un clasificador débil del tipo *Nearest Mean*. Aunque se realizaron pruebas con máquinas de soporte vectorial (SVM), se obtuvieron resultados peores y tiempos de ejecución inadmisibles para el ámbito de la aplicación. Por tanto, una vez obtenido el ángulo de la dirección de cada imagen, en pasos de 30º, se realiza una fusión mediante un filtro Bayesiano. En este último paso se consiguen filtrar reconocimientos erróneos y obtener resultados muy buenos en comparación con las tasas de acierto que obtienen otras técnicas.

En [2] se aborda el problema desde una perspectiva diferente. Basándose en un reconocimiento facial, que ubica la cara gracias al reconocimiento del color de la

[1] Histogram of Shape Contexts (HoSC).

piel en el modelo de color normalizado *rg-chroma*, se reconocen las pupilas detectando el efecto de ojos rojos que causa unos niveles de brillo elevados en la zona correspondiente. Una vez obtenido el plano correspondiente a la cara y la ubicación relativa de las pupilas en ésta, se obtiene la dirección de la mirada, monitorizando así la atención del usuario hacia el lugar en el que se ubica la cámara. Mediante esta técnica se alcanza una tasa de acierto de hasta el 90 %. Sin embargo, en esta tasa de acierto se asume que se reconocen correctamente las pupilas, lo cual solo sucede en un 72 % de los casos como máximo.

Como hemos visto, el problema en cuestión es abordable desde distintos enfoques que conllevan restricciones diferentes, ya sea el uso de múltiples cámaras, cámaras estéreo que permiten capturar una imagen 3D o el uso de otros tipos de sensores. Existen pocos trabajos que empleen exclusivamente una imagen obtenida con una cámara de vista aérea. Destaca la labor de Ozturk et al. en [11] donde se obtiene la dirección de la postura en pasos de 22,5º empleando una imagen capturada en recintos interiores y amplios, como centros comerciales y aeropuertos. En el año 2009, los investigadores afirmaron ser los primeros en abordar el problema de la detección de la postura basándose únicamente en datos de vídeo 2D de sólo una cámara. En concreto, emplean un filtro de partículas a dos niveles, constituido por histogramas de color y de aristas, para realizar el *tracking* del individuo. Para determinar la dirección de la postura se emplean descriptores de forma y puntos SIFT[2]. Mediante estas características se empareja la silueta de la parte superior del cuerpo con una de 16 plantillas previamente definidas. La plantilla que más se asimile al patrón de test determina la dirección reconocida. De esta forma, alcanzan una tasa de acierto del 80 %, dato que está condicionado a obtener una silueta adecuada de la parte superior de la cámara que se parezca a la de entrenamiento. Por tanto, no se han resuelto los problemas de oclusiones o ropa y accesorios que pueda llevar la persona, además de la estatura natural de la misma, que pudieran hacer diferir la silueta de prueba de la de entrenamiento.

Recientemente han aparecido cada vez más técnicas de reconocimiento basadas en inteligencia artificial que hacen uso del llamado *bag of features*. Este conjunto de características consiste en todo tipo de características locales y globales de la imagen y de sus propiedades de color y forma. Esta técnica es empleada en [12], donde el objetivo es determinar la dirección de la postura de los jugadores de un partido de fútbol. Este trabajo forma parte de un proyecto mayor cuyo objetivo final es poder reconstruir en tres dimensiones y en tiempo real un partido de fútbol, lo que permitiría al espectador ver el partido desde cualquier ángulo. Para ello hacen uso de 4 cámaras de las que obtienen la silueta del cuerpo completo de los jugadores a una resolución de entre solo 10x10 a 20x20 píxeles. De estas imágenes toman hasta 18 características distintas, entre las que encontramos las siguientes basadas en la propia imagen: histograma de los niveles de gris, transformaciones de la imagen original (mediante convolución con kernel Gaussiano y kernel Sobel, filtro Gabor, detector de aristas Canny, etc.) y representación de la imagen en el modelo de color LUV. Además se hace uso de otras características basadas en información contex-

[2] Puntos salientes del tipo Scale-Invariant Feature Transform.

tual: posición del jugador, distancia respecto a los otros jugadores, ángulo de los otros jugadores, ángulo de la cámara, ángulo que tenía el jugador en el fotograma anterior y la relación de éste con los ángulos de los jugadores del mismo equipo y del contrario, velocidad de los jugadores, ángulos y distancias a las porterías e histogramas de los ángulos de todos los jugadores, de los contrincantes y del propio equipo. Como vemos, se trata de un amplio conjunto de características de índole muy diferente y cuyas restricciones y utilidades son discutibles y están sujetas a las características de la imagen de test.

Conscientes de esto, no se hace una selección a ciegas, sino que se emplea una selección hacia delante. Este método consiste en un algoritmo voraz que determina la característica óptima localmente para ir seleccionado el vector de características adecuado. Por tanto, se utiliza un conjunto de validación que permite realizar un entrenamiento supervisado y mediante el cual se elige la característica que mejor resultado obtiene en dicho conjunto conocido. Cada característica se utiliza en un clasificador individual del tipo SVM y éstos son combinados mediante la técnica de *Boosting*. Mientras que empleando todas las características se obtiene una tasa de acierto del 62 %, realizando una selección hacia delante la tasa de error es inferior al 1 %. Además, puede observarse que la información contextual es la que más enriquece el sistema al reducir la tasa de error en un 10 %.

Sin embargo, ninguna de estas soluciones logra resolver el problema partiendo únicamente de los datos capturados de una cámara de tipo cenital, que sólo captura imágenes desde el techo de un lugar interior. Por este motivo, se nos plantea un problema nuevo, donde las soluciones existentes no logran su objetivo al disponer de unos datos diferentes. La dificultad intrínseca del problema reside en que se parte de una imagen donde no vemos la cara de la persona y, por tanto, nos tenemos que basar en la postura de la cabeza o del cuerpo.

4.2.2. Clasificación y comparativa de los métodos de resolución actuales

Como se ha comentado en el apartado anterior, existen diferentes formas de abordar el problema de la detección de postura, ya sea considerando distintas fuentes de datos, como imágenes de varias cámaras, imágenes 3D, imágenes frontales, imágenes de vista aérea, otro tipo de sensores, etc., o considerando distintas técnicas de resolución. Entre estas últimas se han visto en el apartado anterior métodos basados en las siguientes estrategias:

- **Plantillas** que se obtienen de las imágenes de entrenamiento y son etiquetadas con una clase discreta para posteriormente emparejarlas con la imagen de test.
- **Clasificadores individuales** que son entrenados para detectar una clase de reconocimiento concreta.
- **Localización geométrica** de los ojos, la boca o la nariz para determinar la dirección de la mirada relacionando la posición de los mismos.

- **Tracking** para localizar el individuo y obtener el cambio global de la dirección de la postura.
- **Métodos híbridos** que combinan éstas y otras estrategias para mejorar las soluciones existentes y sobrepasar sus limitaciones.

Como se ha podido concluir en [13], los métodos basados en plantillas ofrecen la ventaja de poder ampliar fácilmente el conjunto de entrenamiento y las clases de reconocimiento sin tener que realizar cambios relevantes en el sistema. Se trata también de la solución más directa, ya que construir la base de datos de entrenamiento consiste en recortar imágenes de la cabeza o del cuerpo de la persona y realizar un etiquetado manual. No obstante, presentan las desventajas de ofrecer únicamente resultados discretos y que asumen haber obtenido correctamente la región de la cabeza. Finalmente, también dan por hecho que una similitud a nivel de imagen implica una similitud en cuanto a la dirección de la postura. Sin embargo, esto no siempre es así, ya que la apariencia física podría tener más relevancia que la coincidencia de la dirección de la mirada. Este problema se intenta solventar usando características que sean lo más independientes posibles de la apariencia física de la persona y caractericen sólo la postura de la misma, como pueden ser histogramas de la forma de la silueta del BLOB o características obtenidas después de aplicar una transformación a la imagen.

En cuanto a los clasificadores individuales, su principal ventaja consiste en que no es necesario realizar primero una detección de la cabeza o del rostro para posteriormente reconocer la postura, ya que éstos son capaces de distinguir entre cabeza y cuerpo o fondo. Esto se debe justamente a su principal desventaja, y es que para reconocer correctamente una clase deben entrenarse con datos positivos y negativos, por lo que debemos disponer de un conjunto de entrenamiento mucho mayor y el tiempo de aprendizaje puede volverse pesado, sobre todo cuando debemos disponer de un número elevado de clasificadores.

La localización geométrica de las distintas partes de la cara presenta la ventaja de utilizar aquellos elementos que también utilizamos los humanos para poder deducir dónde se dirige la mirada de una persona y, por tanto, se presenta como la estrategia más natural. La mayor desventaja es lógicamente que depende de forma absoluta de la calidad de la detección de las partes a utilizar, por lo que es un método muy afectado por posibles oclusiones o errores de localización.

Los métodos basados en *tracking* realizan un seguimiento de la persona para conocer en todo momento dónde se encuentra ésta en la imagen. Esto conlleva una gran precisión en la fase de localización, pero implica también, en la inmensa mayoría de los casos, utilizar una sincronización inicial o la suposición de empezar el proceso en una dirección determinada. Además, cabe darse cuenta de que, si estos métodos no contemplan una fase adicional de estimación de postura, sólo se obtiene la transformación relativa y no se está detectando la postura de una forma absoluta, lo cual limita las posibles aplicaciones del sistema.

Podemos concluir, por tanto, que se ha realizado una gran labor en los últimos 15 años para poder determinar con cierta exactitud la dirección de la postura o de la mirada de las personas. Existen todo tipo de aplicaciones tanto para espacios interiores como exteriores, y éstas son las que determinan las posibilidades que

tiene el sistema de emplear diferentes tipos de cámaras o de enriquecer el sistema con otro tipo de sensores, además de la visión. Aunque se está lejos de disponer de una solución definitiva que no haga suposiciones y sea lo suficientemente robusta, se ha podido avanzar considerablemente para poder utilizar los sistemas descritos con éxito en casos particulares y bajo condiciones conocidas de antemano.

4.3. Propuesta y desarrollo de un método de resolución

En este apartado proponemos y documentamos el desarrollo de un método de resolución que permite resolver el problema expuesto considerando tres objetivos claramente diferenciables:

1. Detectar y aislar el rostro de una persona mediante técnicas de extracción de movimiento.
2. Determinar la parte del rostro que corresponde a la cabeza del individuo.
3. Reconocer la dirección de la postura de la persona con respecto a la vivienda.

4.3.1. Detectar y aislar el rostro de la persona

Nuestro primer objetivo es detectar los objetos móviles de la escena para poder aislar el rostro de la persona del fondo de la imagen. Para ello existen dos técnicas básicas que cumplen este objetivo. La primera emplea un fondo estático que consiste en una imagen tomada de la escena en cuestión en la que no aparece ningún individuo. Esta imagen de fondo se sustrae de la imagen recibida empleando un margen previamente establecido, diferencia a partir de la cual se considera que el píxel dado es lo suficientemente diferente del fondo como para considerarlo parte del BLOB[3].

Esta técnica tiene la ventaja de ser la más sencilla. Sin embargo, dado que empleamos una única imagen de referencia, todo lo que no esté en esa imagen no será considerado como fondo, por lo que si moviéramos, por ejemplo, el sillón, éste podría ser detectado como parte del BLOB. Por este motivo, hacemos uso de una técnica más robusta que consiste en una extracción dinámica del fondo. En este método se establece una imagen de referencia que va cambiando conforme recibimos más imágenes de la cámara. Es decir, la imagen del fondo se acerca en cada iteración a la imagen nueva, de forma que tras suficientes iteraciones sin cambios, la imagen nueva pasa a ser la imagen de fondo. De este modo, cambios en la habitación, como el movimiento de muebles o la presencia de varias personas, pueden ser considerados al pasar estos cambios a formar parte del fondo una vez pasado un intervalo de tiempo establecido. Existen varios métodos para realizar los acercamien-

[3] Puntos o regiones de la imagen que destacan con respecto al fondo. También conocido como *foreground*

tos entre la imagen de referencia y la imagen nueva, utilizando distintos modelos de color y aproximaciones estadísticas basadas en histogramas o funciones gaussianas, entre otras [14]. Existen también aproximaciones más avanzadas que intentan filtrar las sombras, que al ser también partes móviles de la imagen pueden llegar a considerarse parte del BLOB, así como ofrecer cierta robustez ante cambios de iluminación, como los que aparecen constantemente en las luces fluorescentes [15]. No obstante, dado que éste no es el apartado objetivo de nuestra investigación, nos es suficiente con emplear las soluciones existentes. Finalmente, una vez extraídos los BLOBs de la imagen de esta forma, disponemos de los objetos móviles de la escena, considerando como móviles aquellos objetos que no se han mantenido en una posición fija durante un periodo de actualización de fondo completo. De entre todas las regiones detectadas elegimos la de mayor área, correspondiendo ésta al objeto móvil más grande de la escena y sobre el cual se continuará con el proceso de reconocimiento. Cabe mencionar, sin embargo, que el sistema es fácilmente ampliable para detectar la postura de varias personas, ya que una vez resuelto el método de clasificación sólo habría que ejecutarlo sobre todos los BLOBs que sobrepasen un tamaño mínimo para considerarles individuos, y detectar así la dirección de la postura de todos ellos en una ejecución paralela.

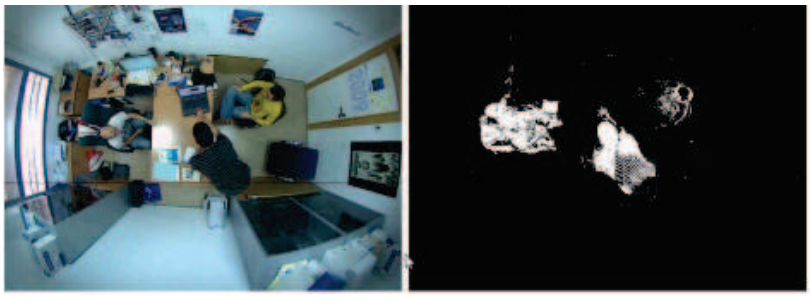

Figura 4.1 Detección de los objetos móviles de la imagen con extracción de fondo dinámica.

4.3.2. Aislar la cabeza del individuo

Una vez determinado el BLOB sobre el cual se debe realizar el reconocimiento cabe estudiar cómo se aborda la clasificación de la imagen para determinar la dirección de la postura. Hay varias soluciones en las que se opta por calcular la elipse que engloba al BLOB, para posteriormente extraer conclusiones a partir de la dirección del eje mayor y del eje menor de la elipse, así como a partir de la relación de tamaño existente entre los ejes [16]. Éste método permite detectar, por ejemplo, si el individuo está de pie, sentado o tumbado, por lo que es capaz de detectar posibles caídas. Sin embargo, aunque se aporta información sobre la postura, el análisis de

la elipse no permite extraer conclusiones sobre la dirección en la que está enfocado el cuerpo.

Otro método que se estudió fue el utilizar los hombros para detectar el eje del cuerpo, suponiendo que éste se encuentra en una postura suficientemente recta y perpendicular al plano. Para lograr este objetivo se empleó el eje menor de la elipse. No obstante, la precisión de este método no era suficiente y disponíamos de una limitación intrínseca de la técnica. Aunque se pudiera obtener correctamente la dirección de la postura del individuo con este método, no seríamos capaces de obtener el sentido de dicha dirección utilizando exclusivamente el eje que conecta los hombros de la persona, y este dato es imprescindible para nuestro reconocimiento.

Así pues, se llegó a la conclusión de que la mayor información sobre la dirección de la postura de la persona se puede deducir a partir de su cabeza. En [7], se obtuvo que la orientación de la cabeza coincide con la mirada en un 87 % de los casos y la contribución que nos da la orientación de la cabeza con respecto a la mirada es de un 69 % de media, llegando hasta un 97 % de tasa máxima.

Inicialmente, se podría pensar que una cabeza vista desde arriba es invariante a rotación si es lo suficientemente redonda y que, en todo caso, nos pasaría igual que con los hombros, pues sólo dispondríamos de la dirección y no del sentido. Sin embargo, no disponemos de una imagen completamente perpendicular al plano, ya que la imagen se toma desde el centro de la habitación y se obtiene una imagen deformada con un efecto similar a un ojo de pez. Esta imagen nos muestra parte de la cara de la persona, por lo que la imagen de la cabeza varía con respecto a las direcciones y al sentido en el que se enfoca la postura del individuo.

Figura 4.2 Imágenes tomadas con una cámara cenital.

Para poder detectar de forma dinámica dónde se encuentra la cabeza del individuo partiendo del BLOB obtenido previamente, se ha aprovechado el tipo de imagen con el que estamos trabajando. Si nos fijamos en una secuencia de imágenes tomadas con la cámara cenital, vemos que siempre se ve parte del cuerpo del individuo en movimiento, y dada la naturaleza radial del punto focal desde el que se toma la imagen, los pies del individuo siempre se encuentran en la posición más cercana al centro, mientras que la cabeza se encuentra siempre a la mayor distancia. Además, estos puntos se pueden obtener de una forma muy eficiente, ya que únicamente cabe evaluar aquellos puntos que definen la frontera entre fondo y BLOB. A continuación, vemos en azul el punto más cercano y el más alejado del centro de la imagen.

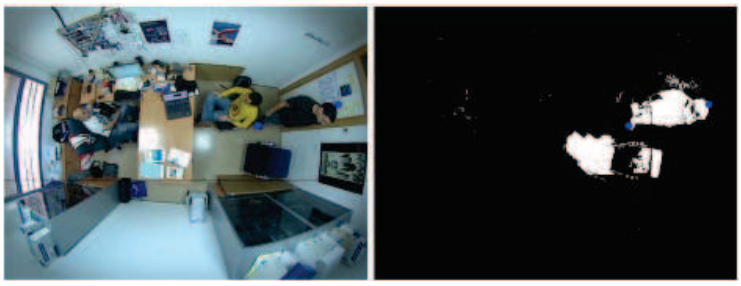

Figura 4.3 Detección del punto más cercano y más alejado del centro de la imagen.

región en la que encontramos la cabeza del individuo. Dado que en la extracción de fondo también partíamos de objetos móviles en la escena, tomamos como aceptable la restricción de que el individuo debe estar de pie.

Existe una excepción a la conclusión expuesta que se da cuando la persona se encuentra justamente debajo de la cámara. Sin embargo, es una situación fácilmente controlable, tanto por la localización de los puntos como por la distancia entre ellos.

Una vez encontrado un punto perteneciente a la región en la que ubicamos la cabeza, podemos encontrar fácilmente la parte correspondiente a la misma. Para ello calculamos una ventana relativa y dinámica, cuya proporción decrece linealmente con respecto a la distancia entre los puntos. De esta forma, la ventana de detección se adapta a la altura de la persona gracias a su definición relativa y, al mismo tiempo, a la posición del individuo dentro de la habitación, gracias a su cálculo dinámico dependiente de lo alejado que esté el BLOB del centro de la imagen. De forma experimental se ha establecido una distancia base, valor que depende del tamaño de la imagen y que se toma como referencia inicial, para poder iniciar un cálculo dinámico. La actualización del porcentaje correspondiente a la cabeza con respecto a la distancia entre el punto más cercano y el más alejado del centro, es decir, entre los pies y la cabeza, se define como sigue:

$$PorcentajeCabeza = 25 * valorBase / distancia(puntoCercano, puntoLejano) \tag{4.1}$$

Donde se puede observar que para la distancia base se otorga un 25 % del BLOB a la cabeza, pero este porcentaje varía linealmente con respecto a la distancia entre los puntos.

Lógicamente, el recorte también debe tener en cuenta el cuadrante en el que se encuentra la cabeza para colocar correctamente la ventana hacia un lado u otro, y hacia arriba o hacia abajo con respecto al punto más alejado.

De este modo, llegamos a obtener con éxito la parte correspondiente a la cabeza del individuo detectado.

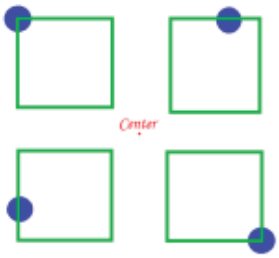

Figura 4.4 Cálculo de la localización correcta de la ventana con respecto al cuadrante.

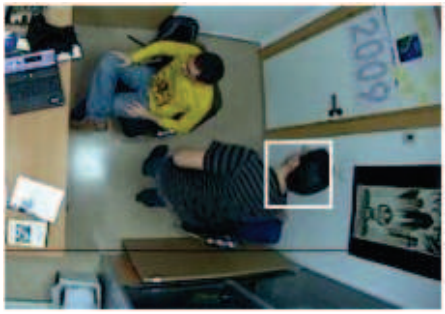

Figura 4.5 Detección de la cabeza del BLOB.

Finalmente, y para cumplimentar nuestro segundo objetivo de aislar correctamente la cabeza de la persona, debemos segmentar la imagen separando el fondo del primer plano.

De nuevo, la idea más sencilla, que consistiría en eliminar de la imagen aquellos píxeles que no se hayan detectado como parte del BLOB, no da un resultado satisfactorio. Esto se debe a que muchos píxeles se detectan como móviles de forma errónea por variaciones de iluminación y porque, debido al *thresholding*[4] aplicado, el contorno del BLOB no es homogéneo, sino que presenta picos, agujeros y esquinas de forma abundante. Por este motivo, se ha optado por realizar un filtrado de fondo que aplique un procesamiento previo a la imagen binaria del BLOB.

Con el objetivo de obtener unos bordes más suaves y homogéneos, se ha aplicado un filtro de emborronamiento gaussiano a la imagen binaria.

Para eliminar las deformidades de la imagen se ha hecho uso de los operadores morfológicos de dilatación y erosión. La dilatación de la imagen provoca que los objetos se expandan, cerrando los agujeros internos del BLOB, mientras que la erosión hace que los objetos se contraigan, de forma que vuelven a su tamaño original, habiéndose eliminado los agujeros en este proceso. Como elemento estructurante se ha utilizado una matriz de 11x11 elementos cuyos valores forman un patrón circular.

[4] Tono, saturación y luminancia.

Cabe tener en cuenta en este punto que estamos tratando con una imagen binaria de aproximadamente 50x50 píxeles, por lo que el coste de realizar estos procesamientos no supone un problema.

A continuación, se muestran los resultados que se obtienen y con los que se abordará el tercer objetivo.

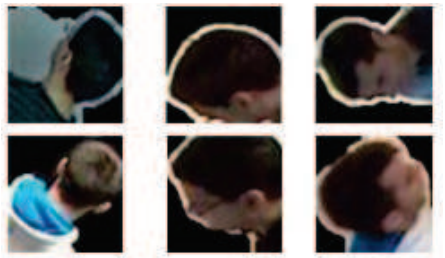

Figura 4.6 Imágenes segmentadas.

4.3.3. Reconocer la dirección de la postura

Partiendo del resultado del punto anterior, nos es posible construir una base de datos de imágenes de distintas personas con diferentes alturas, en habitaciones diferentes y en condiciones de iluminación diversas. Estas imágenes del recorte de la cabeza con el fondo filtrado se etiquetan de forma manual para posibilitar un aprendizaje supervisado.

El objetivo de este tercer apartado es clasificar la imagen comparándola con los datos de referencia, tras haber realizado un aprendizaje previo, y devolver una de entre 8 direcciones posibles.

Figura 4.7 Direcciones a detectar en ángulos de 45º.

4.3.3.1. Mediante análisis de componentes principales

La clasificación de imágenes mediante el análisis de componentes principales, más conocido como PCA, es un método que se basa en obtener unas variables no correlacionadas, denominadas componentes principales, a partir de varias observaciones de una misma fuente, estén éstas relacionadas o no entre sí. Se trata, por tanto, de un método de reducción de dimensionalidad. En el caso de imágenes, se aplica a escala de grises y consiste en descomponer los autovalores de la matriz de covarianza para buscar la transformación que obtiene los componentes principales a partir de los datos de origen. El resultado, aunque disminuido en tamaño y singularidad, mantiene los datos que diferenciaban una muestra de otra y supone que éstos son suficientes para reconocer la clase de la muestra. El emparejamiento se realiza minimizando la distancia Chi-cuadrado de Pearson [20].

4.3.3.2. Mediante varios clasificadores débiles

Si prestamos atención al color de la imagen a clasificar, vemos que, en la mayoría de casos, abundan los colores provenientes del pelo y de la piel. Intuitivamente podría afirmarse que si en la imagen predomina la piel, probablemente el individuo esté mirando hacia la cámara, ya que se ve gran parte de su cara. Sin embargo, si predomina el pelo, es probable que el individuo esté mirando en el sentido contrario, ya que lo estamos viendo desde detrás. Aunque estas afirmaciones son algo vagas, se puede llegar a la conclusión de que según la posición de la persona y dirección de la postura de la misma, la proporción de pelo y piel varía en consecuencia, y esto es un dato que es posible aprender. De aquí proviene la idea de emplear los histogramas de color. Dado que el modelo de color HSV[5] soporta mejor los cambios de iluminación, al variar solo en su valor de luminancia y no modificar sus tres parámetros, como es el caso en RGB, podemos emplear los histogramas correspondientes al tono y la saturación presente en la imagen y descartar los valores del canal de luminancia.

Como hemos visto en el apartado 4.3.2, la imagen que se obtiene una vez filtrado el fondo se corresponde a la cabeza del individuo, y dado que disponemos de su correspondencia binaria, conocemos qué pixeles son cabeza y cuales son fondo. Se nos da por tanto la posibilidad de extraer histogramas de forma. En el caso de imágenes binarias, estos histogramas se dividen en dos instancias: una vertical y una horizontal. Ambas almacenan el número de píxeles a uno que tiene una columna o una fila de píxeles de la imagen. Es por esto que se les denomina histogramas de forma, ya que toman un dato estadístico dependiente de la forma de la imagen. Podemos emplear estos histogramas como datos característicos debido a que según hacia dónde esté mirando la persona la ubicación y la forma de la cabeza en nuestra imagen a clasificar es diferente.

[5] Tono, saturación y luminancia.

El método de clasificación consiste en emplear una fusión de varios clasificadores débiles de tipo KNN[6], donde K define el número de vecinos a contemplar. Cada clasificador KNN trabaja sobre un tipo de histograma concreto. Los resultados de estos clasificadores se contemplan de nuevo como un conjunto de votos del cual se obtiene el valor moda como clasificación final.

Dado que en este caso se están combinando en dos niveles los resultados del cómputo de los vecinos más cercanos para distintos tipos de histogramas, es necesario que contemplemos la distancia concreta a la que se sitúan estos vecinos. Es decir, si durante este proceso obtenemos vecinos muy cercanos, éstos deberían prevalecer sobre otros cuyas distancias respecto a la referencia sean mayores, y emplear únicamente el voto mayoritario desaprovecharía este conocimiento del cual disponemos de por sí al haber computado las distancias entre los histogramas.

Por este motivo se ha contemplado un cálculo del error presente en una votación de os KNN. Para ello, se diferencian los siguientes casos:

- **Primer nivel**: K histogramas de un mismo tipo.
  El error asociado al voto mayoritario se define como:

$$error = | \, Votosdiferentes \, | \, / K \tag{4.2}$$

- **Segundo nivel**: resultados de los cuatro KNN.
  El error asociado al resultado final se define como:

$$error_d = \begin{cases} 4 * error, & \text{si solo hay uno voto} \\ Max(error_i), \forall i, 1 \leq i, \leq 4 / r_i = r_d, & \text{si no} \end{cases} \tag{4.3}$$

  Donde:
  - d identifica la dirección ganadora, es decir, el índice del resultado más votado, y si sólo ha habido un voto, el resultado de menor error de primer nivel.
  - $error_d$ es el error asociado al resultado d.
  - r_i es la dirección asociada al resultado i.

De este modo, es posible obtener una medición del error final que nos indica la calidad del resultado obtenido. Dicho valor puede emplearse, por ejemplo, para desestimar un resultado si el error es mayor que cierto umbral previamente definido. En este caso, el sistema podría no dar ninguna respuesta en lugar de dar una respuesta probablemente errónea y evitar así una experiencia de usuario insatisfactoria. Nos basamos en la premisa de que siempre agrada más al usuario tener que darle al botón dos veces que haber encendido el aparato incorrecto, sin mencionar siquiera aquellos casos donde un resultado erróneo pudiera tener consecuencias indeseadas. Es por ello que disponer de esta medida enriquece al sistema. Además, es posible emplear la medida de error para una fusión de clasificadores, técnica que veremos en el apartado 4.3.3.4

[6] K vecinos más cercanos.

Finalmente, siempre que se empleen clasificadores de tipo KNN debe quedar claro qué tipo de distancia se está empleando. Al disponer de datos numéricos en forma de vector, normalizados a valores entre 0 y 1, disponemos de una gran variedad de medidas de distancia. Éstas presentan distintas complejidades de cómputo y contemplan distintas condiciones de los datos. En nuestro caso, se ha puesto a prueba el método con las siguientes distancias: correlativa, de Bhattacharyya, Chi-cuadrado e intersección.

Basándonos en las pruebas realizadas, la distancia de Bhattacharyya [17] es la que nos proporciona el mejor resultado. Esta distancia se basa en el coeficiente de Bhattacharyya, que indica el grado de coincidencia entre dos distribuciones estadísticas. La principal ventaja de este tipo de distancia es que es proporciona una medida más precisa de la correlación entre dos distribuciones, pero a un coste mayor, ya que computa previamente la suma total de ambas distribuciones.

4.3.3.3. Mediante mapas auto-organizativos

El segundo clasificador que se propone en este trabajo está basado en los mapas auto- organizativos de Kohonen [18]. Se ha optado por los mapas auto-organizativos por dos motivos. En primer lugar, buscábamos un método robusto ante posibles cambios en los patrones de test, los cuales abundan en la aplicación de este trabajo (iluminación, tamaño, lugar, ángulo, etc.). Dadas las vecindades de las características representativas que aprende el mapa, se trata de un método ideal para soportar estos cambios. En segundo lugar, los mapas auto-organizativos permiten aprender los rasgos más característicos de los datos de entrenamiento de forma automática, por lo que aquellas instancias que se repiten en el conjunto de entrenamiento serán representadas en el mapa sin que tengamos que conocer a priori cuáles son las características que queremos aprender. Basándonos en que estas características varían entre clase y clase, el mapa auto-organizativo nos ofrece una forma muy sencilla de clasificar los patrones de test, comparando las coordenadas de las neuronas más similares a estos patrones con las que representan nuestras muestras de entrenamiento.

El siguiente paso es elegir qué características de la imagen nos permiten diferenciar las clases y, además, son adecuadas para emplearse en un mapa auto-organizativo. Es decir, estas características deben ser diferentes entre objetos de clases diferentes.

Así, llegamos a los puntos característicos. Éstos son regiones de la imagen que presentan unas características de cambios de color, iluminación y contraste que destacan sobre el resto de la imagen. Por ello también reciben el nombre de puntos salientes. Estos puntos nos permiten obtener una información local de la imagen que varía entre clase y clase, ya que, en nuestro caso, vemos distintas partes de la cara y de la cabeza según el ángulo en el que esté colocada la persona. Aunque tradicionalmente se han empleado detectores Harris y Susan, éstos se han visto superados en los últimos años por los algoritmos de detección de puntos característicos SIFT, SURF y FAST [21]. Estos últimos sí son invariantes a escala y rotación y son ro-

bustos ante transformaciones y distorsiones de la imagen. Aunque los puntos FAST suponen la opción más sencilla de computar, los puntos SIFT y SURF nos ofrecen una mayor robustez ante cambios de iluminación, escala y transformaciones afines, además de ofrecernos descriptores para codificar la información característica de los puntos. SURF ha demostrado ser mucho más eficiente que SIFT, por lo que se posibilita su uso en sistemas de tiempo real, y supera a SIFT en robustez ante cambios de iluminación [19]. Estos dos factores han sido claves a la hora de decantarnos por los puntos SURF. Cabe comentar, sin embargo, que sus antecesores, los puntos SIFT presentan mayor robustez ante escala y rotación, debido principalmente a un cálculo más preciso y costoso de los ángulos y las escalas de las regiones de la imagen.

Los descriptores que se obtienen a partir de los puntos SURF consisten en vectores de 64 elementos de punto flotante. Estos elementos describen los cambios de nivel de gris que tiene un área de la imagen, tras haber dividido la región característica en 16 zonas, subdivididas a su vez en 4 direcciones, utilizando el ángulo de mayor cambio. De esta forma, obtenemos los descriptores SURF que nos permiten alimentar la red neuronal. En este apartado cabe tener en cuenta qué valores pueden tomar los descriptores para poder inicializar el mapa con neuronas aleatorias, pero factibles.

Figura 4.8 Puntos SURF a clasificar.

Una vez finalizado el entrenamiento del mapa, se ejecuta la fase de etiquetado. En ésta, se obtienen las neuronas correspondientes a cada imagen, calculando sus puntos SURF y emparejándolos con las neuronas más similares.

Para poder clasificar una imagen se han utilizado mapas de calor que se obtienen a partir de las neuronas etiquetadas en la fase anterior. Es decir, cada imagen de entrenamiento dispone de las coordenadas correspondientes a las neuronas que han sido etiquetadas por los puntos de dicha imagen. Colocando el conjunto de coordenadas de cada clase en un mapa diferente, obtenemos los mapas de calor. De esta forma, podemos obtener los puntos SURF de la imagen de test, calcular sus descriptores y emparejarlos con las neuronas del mapa, obteniendo las neuronas que se han activado y que suponen nuestras coordenadas de emparejamiento. Estas coordenadas y sus vecindades permiten obtener una puntuación que indica el grado de coincidencia que han tenido en comparación con el mapa de calor de cada clase. La puntuación máxima se obtiene si existe una coincidencia completa, mientras que una parcial, o una coincidencia de vecindad, obtiene una puntuación linealmente decreciente con respecto a la distancia a la coordenada de referencia más cercana. Evaluando los mapas de cada clase podemos obtener el emparejamiento de mayor

puntuación y, por tanto, mayor correspondencia entre coordenadas de referencia y de test.

4.3.3.4. Fusión de los clasificadores

Aunque en el apartado 4.4 se expondrán los resultados individuales de cada clasificador, nuestro objetivo final es fusionar los resultados de los tres clasificadores. Esto se debe a que se era consciente desde el inicio de que los clasificadores utilizados funcionarían bien en algunos casos y peor en otros. PCA es un método muy eficaz, pero muy sensible ante transformaciones afines, cambios de escala o cambios de iluminación. Nuestro clasificador basado en histogramas soporta muy bien los cambios de iluminación, y el mapa auto-organizativo con puntos SURF es en teoría el más robusto ante cambios de escala y transformaciones afines. Por este motivo, sería ideal poder conocer de antemano las condiciones de la imagen y elegir en función de éstas el clasificador a utilizar. No obstante, no debemos olvidarnos de los parámetros de error o distancia que devuelven nuestros clasificadores. En el caso de PCA disponemos de la distancia al emparejamiento devuelto; en el clasificador basado en histogramas se ha implementado un cálculo propio basado en los resultados de las votaciones KNN[7]; y el clasificador basado en mapas auto-organizativos nos devuelve la puntuación que ha obtenido el mejor emparejamiento.

Basándonos en esta información y en la experimentación realizada, se ha optado por las siguientes reglas:

$$\begin{array}{l} \text{Si distancia}_{PCA} \leq 1000 \text{ entonces retorno } = \text{dirección}_{PCA} \\ \text{Si no Si error}_{Hist} \leq 0.4 \text{ entonces retorno = direccion}_{Hist} \\ \qquad \text{Si no retorno = media (dirección}_{PCA}\text{, dirección}_{Hist}\text{, dirección}_{SOM}\text{)} \end{array} \tag{4.4}$$

De esta forma, optamos por confiar en la dirección obtenida mediante el análisis de componentes principales cuando la distancia del emparejamiento es baja. En caso contrario, utilizamos el clasificador basado en histogramas cuando la decisión mayoritaria obtiene un 60 % o más en el peor de sus emparejamientos. Finalmente, optamos por una decisión conjunta de los tres clasificadores, incluyendo el más complejo, cuando no tenemos suficiente certeza de que los simples nos hayan devuelto un buen emparejamiento.

4.4. Experimentación y pruebas

En este apartado se exponen y evalúan las pruebas que se han realizado sobre el sistema.

[7] Véase el punto 4.3.3.2

Como se ha podido deducir tras la lectura de los apartados anteriores, el sistema desarrollado obtiene tres resultados claramente diferenciables: 1) la detección de la región correspondiente a la cabeza de la persona, 2) la detección de postura de los clasificadores individuales y 3) la detección de postura de la fusión de clasificadores. Por este motivo, se expondrán estos resultados por separado con el objetivo de poder evaluar el funcionamiento de cada una de las fases. Además, se expondrán los resultados del sistema completo en su conjunto, los cuales determinan las tasas de acierto finales del sistema.

En cuanto al entorno de prueba, cabe mencionar que se han ejecutado configuraciones con distintos sujetos y en distintos entornos. Para ello, se ha tenido a disposición la vivienda domotizada del grupo de investigación Domótica y Ambientes Inteligentes de la Universidad de Alicante, denominada *metalTIC - Hogar Digital*. La vivienda tiene tres habitaciones que disponen de una cámara cenital colocada en el centro del techo de cada habitación. Se trata de una cámara Axis 212 PTZ con conexión Ethernet mediante la cual se obtienen un flujo de vídeo a resolución VGA. Para las pruebas se han empleado unas grabaciones de 25 fotogramas por segundo de aproximadamente 10 minutos de duración. El sistema también permite realizar la detección directamente sobre el flujo de vídeo procedente de la casa domótica, pero, como es lógico, las pruebas requieren un etiquetado manual previo, y por ello se han utilizado grabaciones. También cabe mencionar que para conseguir los resultados expuestos ha sido necesario entrenar el sistema con vídeos de entrenamiento procedentes de los mismos entornos y actores.

En cuanto a la implementación y los algoritmos que emplea el método desarrollado, cabe detallar que se ha hecho uso del soporte de las librerías de visión artificial de código abierto OpenCV [22] y Aforge [23].

Vídeo	Actor	Tasa de acierto
2-1	Usuario 1	83,72%
2-2	Usuario 1	84,21%
3-1	Usuario 2	78,69%
3-2	Usuario 2	84,96%
4-1	Usuario 3	80,45%
4-2	Usuario 3	83,65%

Cuadro 4.1 Detección y extracción de la región correspondiente a la cabeza de la persona.

Como puede observarse, se alcanza una media del 82,61 %, que corresponde a la cantidad relativa de fotogramas en los que se consigue detectar y extraer correctamente la región correspondiente a la cabeza de la persona. En esta prueba se han excluido aquellos fotogramas en los cuales no se encuentra la cabeza del individuo en el campo de visión de la cámara. Como vemos, se trata de una tasa de acierto muy satisfactoria, sobre todo si tenemos en cuenta que los fallos se deben, en la mayoría de los casos, a la extracción de fondo dinámica, que se ve superada ante

cambios bruscos de iluminación. Estas dificultades han sido abordadas y superadas en gran parte en otros trabajos [15].

Vídeo	Actor	Clasificador	Tasa de acierto
2-1	Usuario 1	PCA	73,61%
2-1	Usuario 1	Histogramas	30,88%
2-1	Usuario 1	SOM	57,69%
3-1	Usuario 2	PCA	71,91%
3-1	Usuario 2	Histogramas	48,00%
3-1	Usuario 2	SOM	34,00%
4-1	Usuario 3	PCA	52,21%
4-1	Usuario 3	Histogramas	76,59%
4-1	Usuario 3	SOM	35,53%

Cuadro 4.2 Detección de postura de los clasificadores individuales.

En esta prueba se ha evaluado la dirección detectada en pasos de 45º mediante los clasificadores individuales y partiendo de imágenes en las que se realizó correctamente la fase anterior. Es decir, se dispone de un recorte adecuado de la cabeza de la persona. Como se puede observar, el análisis de componentes principales es el de mayor tasa de acierto, alcanzando un 65,91 % de media. Sin embargo, nuestro clasificador débil basado en histogramas llega a superar a este último en algunas pruebas, aunque presenta una tasa media de acierto del 51,82 %. El clasificador basado en mapas auto-organizativos y puntos SURF obtiene con un 42,41 % el peor resultado medio, viéndose superado considerablemente por las opciones iniciales.

Vídeo	Actor	Modalidad	Tasa de acierto
2-1	Usuario 1	1	42,52%
2-1	Usuario 1	2	80,00%
2-2	Usuario 1	1	48,77%
2-2	Usuario 1	2	75,00%
3-1	Usuario 2	1	51,37%
3-1	Usuario 2	2	57,69%
4-1	Usuario 3	1	52,24%
4-1	Usuario 3	2	67,31%

Cuadro 4.3 Detección de postura de la fusión de clasificadores.

Finalmente, llegamos a la evaluación del sistema completo. En esta prueba se han considerado dos modalidades: 1) sin disponer de forma segura de un recorte

adecuado de la región de la cabeza, por lo que en un 17,39 % de casos de media no tenemos una imagen adecuada que clasificar; 2) disponiendo de un recorte adecuado de la cabeza donde la detección y extracción de la región correspondiente ha funcionado correctamente. Es decir, mientras que la modalidad 1 evalúa todo el sistema en su conjunto, la modalidad 2 evalúa solo la clasificación mediante fusión, ya que las tasas de acierto de la detección de la región correspondiente a la cabeza ya han sido analizadas previamente y constituyen una fase independiente del sistema de clasificación.

En la primera modalidad se alcanza una tasa media de acierto del 48,73 %. Este porcentaje equivale a la proporción de fotogramas en los que se detecta correctamente la región de la cabeza del individuo, y dicha región se clasifica adecuadamente obteniendo la dirección de la postura de la persona. No obstante, teniendo en cuenta únicamente la tarea de clasificación de imágenes correspondientes a la región de la cabeza, se obtiene una tasa media de éxito del 70 %.

En cuanto a los tiempos de ejecución del sistema, el análisis completo de las tres fases sobre un fotograma de 640x480 píxeles emplea como máximo 200 ms de ejecución sobre una máquina Core2Duo a 2.2GHz e implementado sobre la plataforma .NET. Por tanto, podemos afirmar que su despliegue en un entorno real es totalmente factible. De este modo, abordando las optimizaciones algorítmicas posibles y empleando el hardware adecuado, se posibilitaría su ejecución a frecuencia de vídeo, lo que es necesario para su uso como interfaz de control.

4.5. Conclusiones

En este trabajo se han abordado varias tareas que han implicado los campos de la visión y la inteligencia artificial en el pre y post-procesamiento de imágenes y flujo de vídeo, así como en la clasificación y el reconocimiento de estos datos. Ha sido necesario un amplio estudio del estado de la cuestión y de las técnicas de solución actuales para poder construir sobre las mismas y abordar el problema conociendo las soluciones existentes y sus deficiencias, pero aprovechando las conclusiones de otros investigadores y proyectos.

La detección de la región correspondiente a la cabeza es uno de los principales aportes de este trabajo, ya que ofrece un sistema robusto y sencillo de computar que obtiene unas tasas de éxito muy satisfactorias por encima del 80 %. En esta tarea cabe destacar también la calidad del recorte que se obtiene de la cabeza del individuo, cuya extracción de fondo y fases de post-procesamiento permiten obtener un resultado útil para todo tipo de aplicaciones de vida asistida por el entorno y vigilancia.

En cuanto a la detección de la postura se ha logrado el principal objetivo: superar la tasa de acierto del análisis de componentes principales. Su tasa del 65,91 % se ve superada por los resultados de la fusión de clasificadores que de media son de un 70 %. Esta mejora se debe principalmente al clasificador débil basado en histogramas y al método de fusión basado en reglas, que logra determinar cuándo se debe

optar por el resultado de un clasificador o de otro. No obstante, los resultados del clasificador basado en mapas auto-organizativos y puntos SURF no son los que se esperaban e indican que debe revisarse las características a emplear y la propiedad de localidad de las mismas, ya que las características globales a la imagen han dado mejores resultados.

Por tanto, las líneas futuras de este trabajo se centrarían en mejorar cada una de las fases del método propuesto, con el fin de aumentar la calidad de la imagen que deben reconocer los clasificadores y mejorar el resultado de éstos. También se nos plantea la posibilidad de exigir algunas suposiciones que puedan mejorar los resultados del sistema, ya que una interfaz de control domótico exige unas tasas de acierto muy altas, porque de esto depende la satisfacción del usuario y la amigabilidad del sistema. Entre éstas encontramos la opción de realizar una sincronización inicial en la que se conoce en qué dirección está mirando la persona, para posteriormente aceptar únicamente cambios de dirección coherentes, basándonos en que, normalmente, no suelen existir giros bruscos de la persona entre un fotograma y el siguiente. Finalmente, se ofrece la opción de integrar este desarrollo en el sistema de servicios de control domótico para el hogar digital del que dispone *metalTIC*.

Referencias

1. A. van Dam: Post-WIMP User Interfaces, In Communications of the ACM, 40(2), pp. 63-67. (1997)
2. F. Wallhoff, M. Ablaβmeier , G. Rigoll: Multimodal Face Detecion, Head Orientation and Eye Gaze Tracking, in Proc. Of Intl. Conf. On Multisensor Fusion and Integration for Intelligent Systems, IEEE. (2006)
3. K. Nickel, R. Stiefelhagen: Real-time Person Tracking and Pointing Gesture Recognition for Human-Robot Interaction, in Proc. of Computer Vision in Human-Computer Interaction ECCV 2004 Workshop on HCI, Czech Republic. (2004)
4. H. Shimizu, T. Poggio: Direction Estimation of Pedestrian from Images, in Proc. of Intelligent Vehicles Symposium, IEEE. (2004)
5. L. Snidaro, C. Micheloni, C Chiavedale: Video Security for Ambient Intelligence, in Proc. of IEEE Transactions On Systems, Man, And Cybernetics - Part A: Systems And Humans, Vol. 35, No. 1. (2005)
6. Q. Jia and X. Yangb: Real-Time Eye, Gaze, and Face Pose Tracking for Monitoring Driver Vigilance, in Proc. of Real-Time Imaging Vol. 8, Issue 5, pp. 357-377. (2002)
7. R. StiefelHagen, J. Zhu: Head Orientation and Gaze Direction in Meetings, In Conference on Human Factors in Computing Systems. (2002)
8. B. Brumitt, J. Krumm, B. Meyers, and S. Shafer: Let There Be Light: Comparing Interfaces for Homes of the Future, IEEE Personal Communications. (2000)
9. P. P. Maglio, T. Matlock, C. S. Campbel, S. Zhai, and B. A. Smith: Gaze and speech in attentive user interfaces, in Proc. of the Intl. Conf. on Multimodal Interfaces. (2000)
10. L. Rybok, M. Voit, H. K. Ekenel, R. Stiefelhagen: Multi-view Based Estimation of Human Upper-Body Orientation, in Proc. of Intl. Conf. on Pattern Recognition. (2010)
11. O. Ozturk, T. Yamasaki, K. Aizawa: Tracking of Humans and Estimation of Body/Head Orientation from Top-view Single Camera for Visual Focus of Attention Analysis, in Proc. of IEEE 12th Intl. Conf. on Computer Vision Workshops. (2009)
12. A. Launila, J. Sullivan: Contextual Features for Head Pose Estimation in Football Games, in Proc. of IEEE Intl. Conf. on Pattern Recognition. (2010)

13. E. Murphy-Chutorian, M. M. Trivedi: Head Pose Estimation in Computer Vision: A Survey, in Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). (2008)
14. G. Bradski, A. Kaehler: Learning OpenCV, Computer Vision with the OpenCV Library, O'Reilly. (2008)
15. D. Grest, J. M. Frahm, R. Koch: A Color Similarity Measure for Robust Shadow Removal in Real-Time, Vision, Modelling and Visualization, pp. 253-260. (2003)
16. H. Nait-Charif, S. J. McKenna: Activity Summarisation and Fall Detection in a Supportive Home Environment, in Proc. of the 17th Intl. Conference on Pattern Recognition, IEEE. (2004)
17. T. Kailath: The Divergence and Bhattacharyya Distance Measures in Signal Selection, IEEE Transactions on Communication Technology. (1967)
18. T. Kohonen: Self-Organizing Maps, Springer-Verlag, Berlin-Heidelberg-New York. (1995)
19. L. Juan, O. Gwun: A comparison of SIFT, PCA-SIFT and SURF, in Proc. of Intl. Journal of Image Processing, Vol. 3, Issue 4. (2010)
20. I. T. Jolliffe: Principal Component Analysis, 2nd ed. New York, Springer-Verlag. (2002)
21. T. Tuytelaars, K. Mikolajczyk: A Survey on Local Invariant Features, Foundations and Trends in Computer Graphics and Vision, Vol. 1, No. 1, pp. 1-94. (2005).
22. OpenCV, librería de visión por computador, `http://opencv.willowgarage.com/wiki/` (Último acceso el 9 de Mayo de 2011)
23. Aforge, framework de visión por computador e inteligencia artificial, `http://www.aforgenet.com/framework/` (Último acceso el 9 de Mayo de 2011)