

# Reconocimiento robusto de gestos en tiempo real para su uso en interfaces basadas en visión

Cristina Manresa Yee

Directores de Tesis: Javier Varona Gómez y Francisco J. Perales López

Dept. de Matemáticas e Informática  
Universitat de les Illes Balears

Ed. Anselm Turmeda. Crta. Valldemossa km. 7.5  
07122 Palma de Mallorca

{[@uib.es](mailto:cristina.manresa,xavi.varona,paco.perales)}

## Resumen

En este documento se describe de forma resumida el trabajo que se ha realizado hasta ahora y que se presentó como DEA [4,5,6].

En este trabajo se presenta una interfaz basada en visión (VBI) que analiza los gestos de la cara y los movimientos de la cabeza del usuario para utilizarlo como dispositivo de interacción con el ordenador. Es una interfaz que trabaja en tiempo real y como entrada del proceso utiliza las imágenes provenientes de una webcam convencional, por lo que se consigue una interfaz de bajo coste, aunque con el consiguiente esfuerzo de trabajar con una calidad de imagen de webcam. El sistema detecta automáticamente la cara del usuario, hace el seguimiento de ciertos puntos característicos de ésta y reconoce los gestos de los guiños.

En la última sección, se presentarán las vías de continuación de la tesis.

## 1. Introducción

El estudio de nuevas formas de interacción más naturales, intuitivas y no invasivas está siendo un campo de investigación en auge, que pretende obtener nuevos sistemas de comunicación a través del reconocimiento y generación del habla o sonidos, visión por computador, animación y visualización gráfica, entendimiento del lenguaje, dispositivos de fuerza, etc.

La información visual aporta muchos datos a los humanos a la hora de interactuar y comunicarse, por lo que si se captura esta información, ésta puede utilizarse para detectar y reconocer acciones y gestos que se analizarán y utilizarán para interactuar con el ordenador.

Cuando un usuario está sentado frente a un ordenador, y hay una webcam colocada sobre el monitor, se asume que la cara está visible. Por ello sistemas de interacción a través de visión que utilizan la detección y seguimiento de puntos característicos de la cara y reconocimiento de gestos de ésta pueden convertirse en una eficiente herramienta de interacción. La importancia de este tipo de sistemas crece aún más, cuando el usuario se ve impedido en utilizar sus manos o brazos (p.e. esclerosis múltiple o distrofia muscular) y por tanto le es imposible hacer uso de los dispositivos tradicionales de interacción como son el teclado y el ratón.

Como requerimientos mínimos para una fluida interacción, la respuesta del sistema debería ser precisa, robusta y en tiempo real. Además las condiciones de entorno como luz o fondo no deberían ser una limitación al sistema.

## 2. Visión general del sistema

Para alcanzar un sistema amigable y fácil de utilizar, el sistema está dividido en dos fases: inicialización y proceso (Fig. 1). La fase de inicialización es la responsable de extraer las características faciales de la cara del usuario. Para ello localiza la cara del usuario, modela el color de la piel, y detecta las posiciones y propiedades de las zonas características: la región de la nariz para seleccionar puntos característicos para realizar el seguimiento y los ojos para detectar los guiños. Este proceso es totalmente automático y el usuario simplemente se tiene que quedar quieto durante unos segundos para que se lleve a cabo de forma satisfactoria.

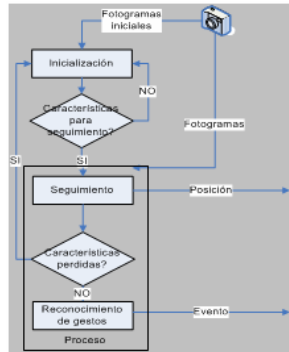


Figura 1. Sistema en dos fases: Inicialización y Proceso

La fase de proceso se encarga de hacer el seguimiento de los puntos característicos encontrados en la región de la nariz a través del algoritmo de Lucas y Kanade, el seguimiento de los ojos a través de distribuciones de color y el reconocimiento de gestos.

Finalmente, toda la información obtenida en la fase de proceso, se transformará en una posición del cursor del ratón y en una ejecución de un evento del ratón si procede. El sistema se reiniciará automáticamente si se pierden los puntos característicos y se re-llamará al módulo de inicialización

### 3. Aprendizaje de las características faciales

Es importante que el sistema sea lo más simple posible para el usuario, por lo que se ha evitado un proceso de calibración donde sea necesario que el usuario interfiera y realice una serie de procesos determinados. Debido a este requerimiento, el sistema detecta automáticamente la cara del usuario a través del algoritmo robusto de Viola y Jones [8] de detección de caras en tiempo real. Cuando el sistema se ejecuta al iniciarse, el usuario se debe quedar estático para que se realice la fase de inicialización. La detección de la cara se considera robusta cuando durante una serie de fotogramas se detecta la cara sin cambios en su posición (Fig. 2(a)), eso significará que el usuario quiere utilizar el sistema. Gracias a las medidas antropométricas de la cara, ésta puede dividirse en tres regiones: ojos y cejas, nariz y boca.

Sobre la zona de la nariz, se buscan los mejores puntos característicos para realizar su seguimiento a través del algoritmo de Shi y

Tomasi [7]. Normalmente se seleccionarían puntos en las fosas nasales y en los bordes de la nariz, pero debido a las condiciones de luz, puede ocurrir que se marquen puntos que no están sobre las posiciones deseadas (Fig. 2(b)). Lo ideal es que los puntos seleccionados estén en ambas partes de la nariz y con ciertas condiciones de simetría con respecto al eje vertical. Por ello, se hará una reelección de los puntos teniendo en cuenta este requerimiento (Fig. 2(c)). Finalmente la Fig. 2(d) ilustra el punto final que se considerará para convertirlo en la posición del cursor. Este punto es una media de todos los puntos característicos del grupo seleccionado.

A continuación, se modela el color de la piel, que será utilizado en el módulo de reconocimiento de gestos para limitar el proceso de los píxeles a aquellos situados dentro de las fronteras de una máscara de la cara. Las muestras para aprender el color de la piel serán tomadas de la región de la cara y la función de densidad se representará con una gaussiana en 3D RGB. Los valores de los parámetros del modelo gaussiano (media y covarianza) están calculados utilizando métodos de aprendizaje estadístico [3]. Una vez calculados estos parámetros, se podrá crear una máscara de la cara para el usuario.

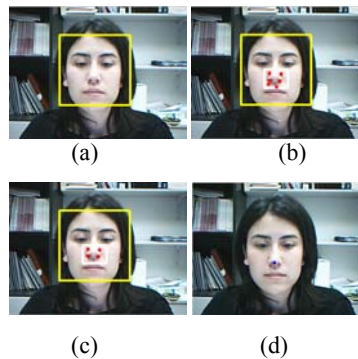


Figura 2. (a) Detección automática de la cara. (b) Conjunto inicial de puntos característicos. (c) Reselección del conjunto de puntos considerando simetrías. (d) Media de los puntos

Finalmente falta la construcción de los modelos de los ojos del usuario. El algoritmo de seguimiento de ojos que se explicará en la sección siguiente se basará en el color del ojo. Éste está compuesto por el iris y la esclerótica, y para

modelarlo se utilizan técnicas de histograma en el espacio de color RGB. En nuestros resultados, los histogramas están calculados utilizando  $16 \times 16 \times 16$  bins. Además se añade una función de ponderación a cada bin dependiendo de su distancia al centro.

#### 4. Seguimiento de características

El seguimiento de los ojos es a través de su distribución de color. Al ponderar el modelo del ojo con un kernel isotrópico hace que sea posible utilizar una función de optimización del gradiente, como el algoritmo mean-shift, para buscar el modelo del ojo en el nuevo fotograma [1]. En nuestro sistema este algoritmo funciona de forma correcta y en tiempo real y aunque pudieran ocurrir pequeños errores de posición, en nuestra aplicación no afectan demasiado ya que lo que se quiere es detectar la zona del ojo para analizar si se produce un guiño. Además siempre se tiene la frontera de que esas regiones tienen que encontrarse dentro de la máscara de la cara.

El seguimiento de los puntos sobre la región de la nariz se realiza a través de la información del gradiente de intensidad espacial [2]. Este algoritmo es robusto ante rotaciones, escalados o deformaciones de la imagen, por lo que el usuario puede moverse de forma libre. Movimientos rápidos del usuario o cambios de iluminación bruscos, puede hacer que se desplacen o pierdan los puntos característicos a seguir, por lo que si un punto característico se distancia una cierta longitud del punto medio calculado, este punto será descartado. Ejemplos de esta fase de seguimiento de nariz y ojos pueden verse en la Fig. 3.

El punto medio de todos estos puntos será el que se transforme después a la posición del cursor por lo que para suavizar los movimientos del ratón y que la trayectoria del cursor no tenga discontinuidades, se utiliza un método de regresión lineal entre fotogramas.

#### 5. Reconocimiento de gestos

Los gestos a reconocer son los guiños, cuya detección es complicada debido a la calidad de las webcams. El proceso depende de la posición de la cara del usuario, por lo que suponemos que estará de frente a la cámara.



Figura 3. Secuencia de imágenes de seguimiento

Basamos nuestra detección en encontrar el contorno del iris, si se encuentra consideraremos que el ojo estará abierto, y por el contrario, si no se detecta, el ojo se considerará cerrado. El proceso se inicia con la detección de los contornos verticales sobre la zona de los ojos. Para evitar falsos positivos, se utilizan operadores lógicos con una máscara del ojo conseguida a través de binarizarla previamente y por último, se mantienen los dos bordes más largos considerando un mínimo de píxeles. Si estos bordes no aparecen a lo largo de una serie de fotogramas, se considerará que el usuario tiene el ojo cerrado. En la Fig. 4 se ve el proceso seguido

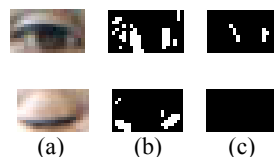


Figura 4. Proceso de reconocimiento de guiños

#### 6. Aplicación y resultados

A partir de los procesos descritos anteriormente, éstos se pueden fusionar para obtener un sistema completo de interacción con el ordenador que reemplace al ratón.

La precisión que se necesita del control de la posición del ratón es aquella que permite al usuario desplazar el cursor a la zona deseada. Esto se hace a través de un mapeo de la posición de la

nariz a la pantalla teniendo en cuenta la posición anterior sobre la pantalla.

El sistema de interacción ha sido probado por un conjunto de 22 usuarios que no presentaban ninguna discapacidad, y de los cuales la mitad no había experimentado nunca con este sistema antes y los demás habían entrenado alrededor de cinco minutos. Se contaba con una plantilla de una parrilla de 5x5 círculos de radio 15 píxeles repartidos por la pantalla. El usuario tenía una única oportunidad de situarse encima de ellos y pulsar. A medida que el usuario realizaba la tarea se guardan datos de distancias de pulsaciones fallidas. En la tabla siguiente se resumen los resultados.

Grupo Usuarios	Clicks reconocidos	Distancia media de errores
Entrenados	97,3 %	2 píxeles
Noveles	85,9 %	5 píxeles

Figura 5. Tabla de resultados de las pruebas

## 7. Conclusiones y trabajo futuro

Como conclusión del trabajo, se ha demostrado que se ha obtenido una interfaz a través de visión por computador de bajo coste y capaz de sustituir el ratón a través de utilizar movimientos de la cara y gestos. Este sistema de interacción está disponible gratuitamente en Internet, y fue el proyecto ganador en el II Premio Fundetec 2006 en la categoría de 'Mejor Proyecto de Entidad No Lucrativa destinado a Ciudadanos'.

Una de las partes más importantes de este sistema son sus usuarios potenciales, y aunque tenga una vertiente de ocio, la e-inclusión y la e-accesibilidad del mayor número de personas dentro del conjunto de usuarios que no pueden hacer uso de los dispositivos tradicionales son las principales metas buscadas, y por tanto base de la tesis.

En estos momentos se está trabajando en un proyecto que tiene como intención hacer una evaluación completa de usuarios de diferentes edades que presentan discapacidades tanto físicas como mentales, para ver que mejoras se podrían realizar y que dificultades tienen los usuarios a la hora de utilizar el sistema.

Una mejora que se puede hacer para que pueda ser utilizado por un sector más grande de

usuarios es hacer la aplicación más robusta en cuanto a la recuperación de la posición si el usuario presenta movimientos espasmódicos.

Con el objetivo de ofrecer una interacción más rica y con más posibilidades se podría obtener una interfaz multimodal, que incluiría un reconocimiento de sonidos o habla (ARS) para generar eventos del ratón u otras acciones que se predefinieran y sería interesante incluir un conjunto más amplio de gestos o combinaciones de éstos a reconocer para acelerar acciones.

Un avance importante sería el reconocer emociones para conocer el estado del usuario como alegría, tristeza, enfado etc. y actuar en consecuencia.

## Referencias

- [1] Comaniciu, D., Ramesh, V., Meer, P.: Kernel based Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (2003) 564–577
- [2] Baker, S., Matthews, I.: Lucas-Kanade 20 Years On: A Unifying Framework. *International Journal of Computer Vision* 56 (2004) 221–225
- [3] Bishop, C.M.: *Neural Networks for Pattern Recognition*. Clarendon Press (1995)
- [4] Manresa-Yee, C., Varona, J., Perales, F.J.: Face-Based Perceptual Interface for Computer-Human interaction. *WSCG'2006 Short Communication Proceedings ISBN 80-86943-05-4*.
- [5] Manresa-Yee, C., Varona, J., Perales, F.J.: Non-verbal communication by means of head tracking. *Ibero-American Symposium on Computer Graphics - SIACG06, 2006 pp. 72-75*
- [6] Manresa-Yee, C., Varona, J., Perales, F.J.: Towards hands-free interfaces based on real-time robust facial gesture recognition *AMDO'06, Lecture Notes in Computer Science 4069: 504-513, 2006*.
- [7] Shi, J., and Tomasi, C.: Good Features to Track. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (1994) 593–600*
- [8] Viola, P., Jones, M.: Robust Real-Time Face Detection. *International Journal of Computer Vision* 57 (2004) 137–154