CAEPIA'99, Vol. 1, Pág. 48-53. Murcia, Nov. 1999.

Seguimiento robusto del rostro humano mediante visión computacional

J.M. Buenaposada L. Baumela
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid
Campus de Montegancedo s/n, 28660 Boadilla del Monte (Madrid)
e-mail: jmbuena@mayor.dia.fi.upm.es, lbaumela@fi.upm.es

Palabras clave: Seguimiento del rostro humano, visión computacional, interfaces multimodales

Resumen. *La construcción de interfaces multimodales robustas basadas en visión computacional es aún un problema abierto. En este artículo presentamos la componente más elemental de una interfaz multimodal basada en visión computacional: un seguidor de caras. Hemos desarrollado un sistema robusto de seguimiento en tiempo real del movimiento del rostro humano. Está basado en una arquitectura de control que coordina a varios algoritmos de seguimiento de características complementarias. Estos algoritmos colaboran dentro de la arquitectura para llevar a cabo un seguimiento robusto. Esta filosofía de funcionamiento permitirá al sistema adaptarse a condiciones de trabajo cambiantes (variaciones de la iluminación, oclusiones, etc.), obteniendo en todo momento la máxima infomación posible sobre la posición y orientación del rostro.*

1 Introducción

La necesidad de construir interfaces más naturales, atractivas y fáciles de utilizar ha dado lugar a la aparición de una línea de trabajo en lo que se conoce como interfaces multimodales [Waibel96]. Esta forma de concebir la comunicación con la computadora se caracteriza por la integración de distintos dispositivos (reconocimiento y síntesis de voz, visión computacional, teclado, ratón, lápiz óptico, etc.) con el objetivo de proporcionar un conjunto alternativo y redundante de vías de comunicación.

El trabajo que presentamos en este artículo, un seguidor del movimiento del rostro humano, es la pieza elemental sobre la cual construir un bloque de interpretación de gestos faciales para una interfaz multimodal. Conociendo de una manera precisa y robusta la posición del rostro, podremos plantearnos el identificar un lenguaje de gestos en base al análisis del movimiento de algunas de sus partes más expresivas (boca, ojos, cejas, etc.).

Los sistemas de seguimiento del rostro humano que se han construido hasta la fecha pueden agruparse en tres categorías: (a) seguimiento 2D, realizan el seguimiento sólo en posición [Isard96, Bradski98]; (b) seguimiento $2\frac{1}{2}$D, realizan un seguimiento 2D con alguna información de orientación [Hager96, Rae98]; (c) seguimiento 3D, realizan el seguimiento en los seis grados de libertad de la cabeza [Gee96, Stiefelhagen97].

La práctica totalidad de los sistemas construidos se caracterizan por utilizar un único modelo de seguimiento. Los que realizan un seguimiento 3D son menos robustos que los que realizan

seguimiento 2D, pero son mucho más precisos y permiten analizar el movimiento de todas las partes de la cara. Por el contrario, los sistemas de seguimiento 2D emplean primitivas muy simples (color, bordes, etc.) y por ello son capaces de trabajar en unas condiciones más adversas (mala iluminación, por ejemplo), permitiendo una fácil recuperación ante fallos de seguimiento.

El sistema que presentamos en este artículo se fundamenta sobre una arquitectura de seguimiento redundante. Está formada por dos "seguidores" basados en modelos de seguimiento diferentes (2D y $2\frac{1}{2}$D) y un autómata de control que coordina su fucionamiento. Esta filosofía de comportamiento permitirá al conjunto responder de una manera flexible a las variaciones de las condiciones de trabajo (iluminación, oclusiones, etc.) obteniendo en cada momento toda la información disponible sobre la posición y orientación del rostro. De ahí que califiquemos al sistema como "robusto". Cuando las condiciones de trabajo sean óptimas, seguirá con precisión el movimiento del rostro. A medida que estas condiciones se deterioren, el modelo de seguimiento será menos preciso. El sistema funcionará de manera dinámica, de modo que en cada momento el autómata seleccionará el modelo de seguimiento más adecuado.

El interés de este trabajo radica en utilizar un modelo de seguimiento redundante que permita aumentar la robustez y flexibilidad de sistema ante condiciones de trabajo cambiantes.

2 El sistema

El objetivo fundamental que se persigue con la construcción del sistema de seguimiento es disponer de una base experimental para la construcción de interfaces multimodales basadas en el análisis de gestos faciales. Por ello el sistema no sólo detecta la presencia de un rostro y sigue su centro de masas en la imagen, sino que, además, cuando mira de frente, es capaz de seguir con gran precisión la posición del rostro y su orientación respecto al eje axial de la cámara, de modo que se pueda calcular con exactitud la posición y orientación de los ojos, boca y cejas.

El sistema está organizado en tres niveles coordinados por un autómata finito (ver figura 1). Cada uno de los niveles se corresponde con un seguidor de características diferentes.

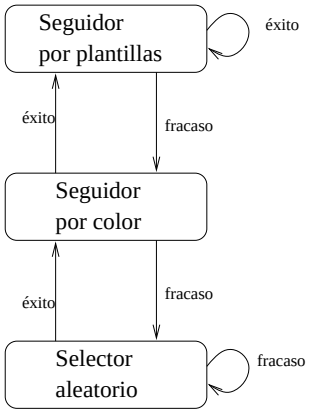

Figura 1: Diagrama de estados del sistema de seguimiento.

En el nivel más bajo se sitúa un selector aleatorio de rectángulos. Este nivel se activa cuando el sistema ha perdido completamente el rostro que estaba siguiendo. Busca aleatoriamente un rostro en el campo de visión de la cámara.

El segundo nivel es un seguidor 2D que detecta grupos de píxeles del color de la piel. El seguidor construido es una versión modificada del algoritmo Camshift [Bradski98] al cual, para mejorar su funcionamiento, se ha añadido una etapa previa basada en un clasificador bayesiano [Buenaposada99]. Este seguidor sigue el movimiento del mayor grupo de píxeles que tengan el color de la piel, calculando al mismo tiempo su eje de mínima energía y su área. Estos datos son indicadores de la posición 3D del rostro y de su orientación en el plano imagen.

El tercer nivel de seguimiento está formado por un seguidor $2\frac{1}{2}$D. Este seguidor, basado en el algoritmo SSD [Hager96], parte de una estimación aproximada de la posición y orientación del rostro, que optimiza mediante un proceso de ajuste mínimo cuadradático a un modelo construido previamente. Para mejorar su eficiencia computacional hemos simplificado el modelo de invarizas a cambios de iluminación, empleando en nuestro sistema una normalización de la imagen basada en la ecualización de su histograma [Buenaposada99].

Cada uno de estos algoritmos de seguimiento está continuamente controlando su propia ejecución. Cuando el seguimiento se realiza de una manera satisfactoria, cede el control a un algoritmo de nivel superior, que realizará el seguimiento con mayor precisión. Cuando detecte un fallo, transitará a un seguidor de nivel inferior, que es más robusto y rápido. En caso de pérdida total, el seguidor de nivel más bajo realizará una búsqueda aleatoria del rostro por toda la imagen.

A continuación describimos en más detalle los fundamentos de cada uno de los algoritmos de seguimiento empleados.

2.1 Seguimiento 2D

El algoritmo de seguimiento 2D está compuesto por dos etapas. Una primera etapa de segmentación que determina, dentro de una ventana de interés en la imagen, qué píxeles tienen un color semejante al de la piel. La segunda etapa, la de seguimiento, calcula la posición y orientación de la región que contiene al mayor grupo de píxeles con el color de la piel, y modificará la posición y orientación de la ventana de interés para que ésta aparezca siempre centrada sobre dicha región.

2.1.1 Segmentación estadística

Estudios cuantitativos de la distribución del color de la piel humana concluyen que las diferencias en la apariencia del color de la piel bajo distintas condiciones de trabajo dependen más de la intensidad que del color en sí mismo [Yang98]. Normalizando, por tanto, el modelo de color RGB respecto a la intensidad (R+G+B) obtenemos el modelo RG normalizado (r_n, g_n), donde $r_n = R/(R+G+B)$ y $g_n = G/(R+G+B)$.

Los píxeles de la piel de la cara ocupan una zona bien delimitada del espacio de color normalizado (ver Figura 2a) que puede modelizarse mediante una fdp normal [Buenaposada99], $p(r_n, g_n|\alpha_p)$, que representa la probabilidad de que en la clase "piel" aparezca un píxel con color (r_n, g_n). Por otra parte, la probabilidad de que dicho píxel se presente en la clase "no

piel", $p(r_n, g_n|\alpha_{np})$, será una función uniforme que ocupe todo el espacio de color, puesto que todo lo que no sea piel tendrá la misma probabilidad de tener un color cualquiera.

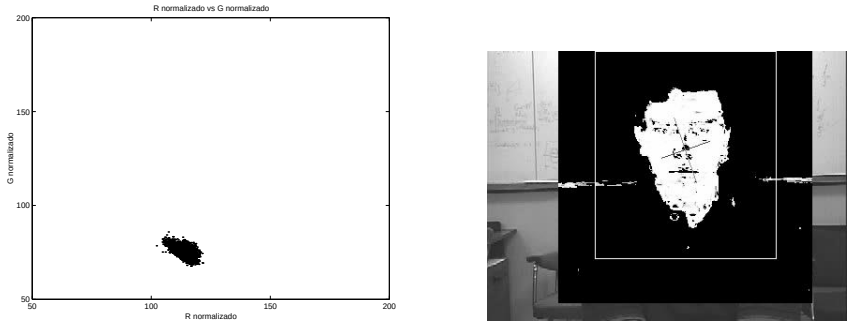

Figura 2: (a) Espacio de color de los píxeles de la piel de la cara. (b) Resultado del proceso de segmentación.

El resultado de esta etapa es una imagen en la que el nivel de gris de cada píxel, $g_{np}(x, y)$, representa su grado de pertenencia a la clase "piel" (ver Figura 2b),

$$g_{np}(x, y) = 255\, p[\alpha_p|r_n(x, y), g_n(x, y)] = 255\, \frac{p[r_n(x, y), g_n(x, y)|\alpha_p]p[\alpha_p]}{p[r_n(x, y), g_n(x, y)]},$$

donde $r_n(x, y), g_n(x, y)$ representa el color del píxel de coordenadas (x, y), $p[r_n(x, y), g_n(x, y)]$ es la probabilidad total de ocurrencia de un píxel con color $r_n(x, y), g_n(x, y)$ y $p[\alpha_p]$ es la probabilidad a priori de la clase "piel".

2.1.2 Seguimiento

Partiendo de $g_{np}(x, y)$, el algoritmo de seguimiento calcula iterativamente el centro de masas del mayor grupo de píxeles con el color del la piel, (x_c, y_c), su eje de mínima energía, θ, que representa la orientación del rostro en el plano imagen, y el tamaño de la ventana de interés, representado por la longitud de uno de sus lados, s:

$$(x_c, y_c) = \left(\frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}}\right);\ s = 2\sqrt{\frac{M_{00}}{255}};\ \theta = \frac{1}{2}\arctan\left(\frac{2b}{a-c}\right),$$

donde M_{ij} representa el momento de orden (ij) de $g_{np}(x, y)$ y

$$a = \frac{M_{20}}{M_{00}} - x_c^2;\ b = 2\left(\frac{M_{11}}{M_{00}} - x_c y_c\right);\ c = \frac{M_{02}}{M_{00}} - y_c^2.$$

2.2 Seguimiento $2\frac{1}{2}$D

Los parámetros de posición y orientación del rostro estimados en la sección 2.1 se optimizan mediante un ajuste minimocuadrático a una plantilla del rostro (ver figura 3).

Sea I el vector en el que se almacenan los n píxeles de la plantilla, e $I(\bar{\mu})$ el resultado de aplicar sobre I una transformación de parámetros $\bar{\mu}$. Un desarrollo en serie de Taylor de primer orden permite establecer la relación

$$I(\bar{\mu} + \delta\bar{\mu}) = I(\bar{\mu}) + \mathbf{M}\delta\bar{\mu}, \tag{1}$$

donde $I(\bar{\mu} + \delta\bar{\mu})$ es la imagen del rostro, $\delta\bar{\mu}$ es el error cometido al estimar $\bar{\mu}$ y $\mathbf{M}$ es el jacobiano de $I(\bar{\mu})$ respecto de $\bar{\mu}$.

Si la estimación inicial de los parámetros de la cara, $\bar{\mu}$, es lo suficientemente precisa, $\delta\bar{\mu}$ se puede obtener resolviendo (1) mediante mínimos cuadrados:

$$\delta\bar{\mu} = \mathbf{M}^{-}\bar{\varepsilon} \tag{2}$$

siendo $[\,]^{-}$ el operador inversa generalizada y $\bar{\varepsilon} = [I(\bar{\mu} + \delta\bar{\mu}) - I(\bar{\mu})]$ la disparidad entre dos instantáneas de la secuencia.

El mayor inconveniente que presenta un algoritmo de seguimiento basado en (2) es la carga computacional de invertir una matriz $\mathbf{M}$ de dimensión $n \times m$ (n es el número de píxeles de la plantilla y m el número de parámetros de $\bar{\mu}$). Sin embargo, eligiendo un modelo de transformación de la plantilla adecuado (p.ej. un modelo afín), $\mathbf{M}$ puede descomponerse en [Hager96]

$$\mathbf{M} = \mathbf{M}_0\mathbf{\Sigma}(\bar{\mu}), \tag{3}$$

donde $\mathbf{M}_0$ es una matriz que depende de la estructura de la plantilla, y por lo tanto es constante, y $\mathbf{\Sigma}(\mu)$ depdende de los parámetros del modelo de transformación elegido.

El seguidor que hemos construido emplea un modelo de transformación de la plantilla basado en cuatro parámetros: un ángulo de rotación sobre el plano imagen, θ, un escalado de magnitud s y una traslación $(\Delta x, \Delta y)$. La descomposición (3) para este modelo sería

$$\mathbf{M}(\mu) = \underbrace{\begin{bmatrix} \nabla I^t_{x_1y_1}\Gamma(x_1, y_1) \\ \cdots \\ \nabla I^t_{x_ny_n}\Gamma(x_n, y_n) \end{bmatrix}}_{\mathbf{M_0}} \underbrace{\begin{bmatrix} \frac{1}{s}\mathbf{R}(-\theta) & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{s} \end{bmatrix}}_{\mathbf{\Sigma}(\bar{\mu})}, \tag{4}$$

donde $\nabla I^t_{x_iy_i}$ es el vector gradiente de la plantilla en el píxel de coordenadas x_i, y_i y

$$\Gamma(x_i, y_i) = \begin{bmatrix} 1 & 0 & -y_i & x_i \\ 0 & 1 & x_i & y_i \end{bmatrix},$$

con lo que el ajuste de los parámetros que modelizan el desplazamiento del rostro vendría dado por

$$\delta\bar{\mu} = \mathbf{\Sigma}^{-1}(\bar{\mu})\mathbf{M}_0^{-}\bar{\varepsilon}$$

en la que $\mathbf{M}_0^{-}$ se calcularía off-line para la plantilla empleada.

Para el cálculo de la disparidad, $\bar{\varepsilon}$, empleamos una pirámide multirresolución a tres niveles (ver figura 3). A medida que la disparidad aumenta disminuimos la resolución con la que se comparan las imágenes, con lo que mejora el comportamiento del sistema frente al "aliasing" y aumenta la velocidad de proceso al trabajar con imágenes más pequeñas.

Asimismo, ecualizamos la imagen y la plantilla antes de compararlas. Este proceso normaliza las distribuciones de los niveles de gris de ambas imágenes y amortigua los cambios producidos por variaciones en la iluminación.

Figura 3: Plantilla del rostro ecualizada a distintas resoluciones.

3 Resultados

En esta sección evaluaremos cuantitativamente el rendimiento del sistema de seguimiento en su conjunto y el de cada uno de sus componentes.

En la figura 4 se muestran gráficas comparativas de la precisión en el seguimiento del rostro en una secuencia compuesta por 80 imágenes. Por simplicidad sólo se muestra el seguimiento en la coordenada horizontal, que es en la que se producen los mayores desplazamientos.

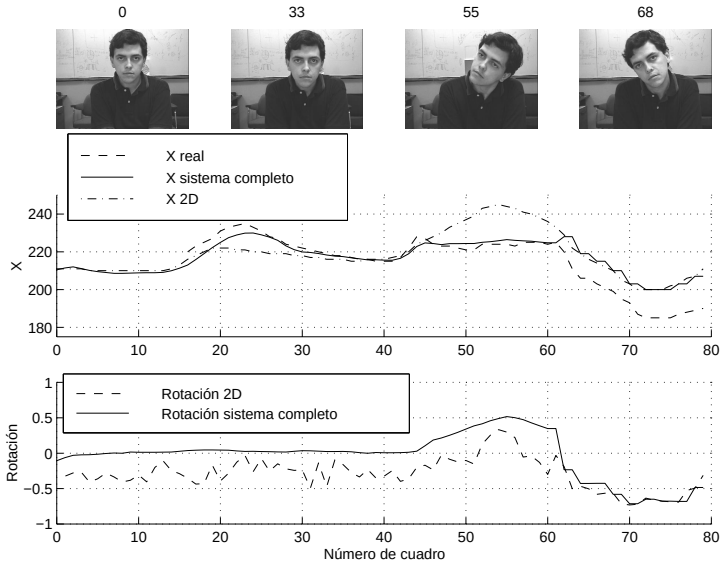

Figura 4: Precisión en el seguimiento de una secuencia.

En la primera gráfica se comparan la coordenada horizontal calculada por el seguidor 2D funcionando aisladamente, la calculada por el sistema de seguimiento completo y la posición real del rostro calculada manualmente. Podemos inferir que el seguidor $2\frac{1}{2}$D aislado es el menos robusto, pues a partir de la imagen 60 se pierde, ya que el error de seguimiento del sistema completo coincide con el del seguidor 2D. Por su parte, el seguidor 2D es mucho más robusto, pues nunca pierde al objetivo, aunque lo sigue con mucha menos precisión que la obtenida por el sistema completo. Del mismo modo, el seguimiento de la rotación del rostro en torno al eje Z oscila mucho menos cuando se realiza el seguimiento con el sistema completo que con el 2D. El seguimiento resultante del sistema completo posee las características más deseables. Cuando las condiciones lo permiten realiza un seguimiento $2\frac{1}{2}$D, que es el más preciso, y cuando éste seguidor se pierde continúa haciendo un seguimiento 2D, que es menos preciso pero más robusto.

En la figura 5 se muestra una secuencia de evaluación del rendimiento del sistema frente a oclusiones. En la gráfica podemos apreciar cómo en cada oclusión se produce una pérdida momentánea del rostro, puesta de manifiesto por un pico en la estimación de la posición horizontal y por un cambio de capa en el algoritmo de seguimiento (en la tercera gráfica de la figura 5 capa 0 representa el selector aleatorio, capa 1 al seguidor basado en color y capa 2 al seguidor basado en plantillas). El sistema se recupera rápidamente cuando el rostro vuelve a aparecer.

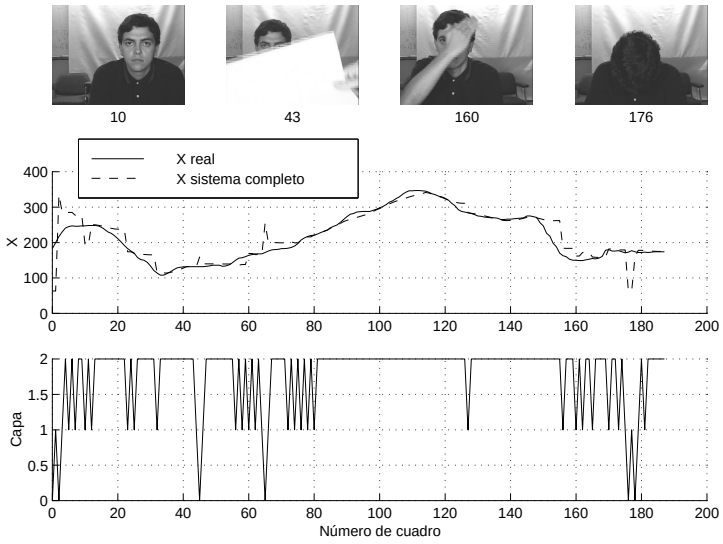

Figura 5: Secuencia con oclusiones

4 Conclusiones

Hemos presentado un sistema robusto de seguimiento del rostro humano basado en la coordinación de dos algoritmos de características complementarias. El sistema construido es capaz de procesar hasta 10 imágenes por segundo con un rendimiento claramente superior al que presentan cada uno de estos algoritmos de seguimiento aisladamente.

Entre las aportaciones cabe mencionar la introducción del proceso de ecualización para mejorar la robustez ante variaciones en la iluminación y la utilización del clasificador estadístico basado en modelo de color RG normalizado para segmentar la imagen.

Las futuras líneas de trabajo irán encaminadas, por un lado, a mejorar el rendimiento del sistema empleando nuevos algoritmos de seguimiento y mecanismos más sofisticados de coordinación que incluyan paralelismo en la ejecución de los seguidores. La otra línea de trabajo consistirá en emplear este primer prototipo de sistema de seguimiento como plataforma para la construcción de un sistema de análisis de gestos faciales.

Referencias

[Bradski98] G.R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Proceedings IEEE Workshop on applications of computer vision* (WACV98), 214-219. 1998.

[Buenaposada99] J.M. Buenaposada. Seguimiento robusto del rostro humano basado en visión artificial. Trabajo Fin de Carrera. Facultad de Informática. Universidad Politécnica de Madrid. Julio 1999.

[Gee96] A. Gee, R. Cipolla. Fast visual tracking by temporal consensus. *Image and Vision Computing* 14(2), 105-114. 1996.

[Hager96] G. Hager, P.N. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. *Proceedings International Conference on Computer Vision and Pattern Recognition*, 403-410. 1996.

[Isard96] M. Isard, A. Blake. Contour tracking by stochastic propagation of conditional density. *Proceedings European Conference on Computer Vision* (ECCV96). Vol. I. 343-346. Abril 1996.

[Rae98] R. Rae, H.J. Ritter. Recognition of human face orientation based on neural networks. *IEEE Transactions on Neural Networks*, 9(2) 257-265. 1998.

[Stiefelhagen97] R. Stiefelhagen97, J. Yang, A. Waibel. A model-based gaze tracking system. *International Journal on Artificial Intelligence Tools*, 6(2) 193-209. 1997.

[Waibel96] A. Waibel, M.T. Vo, P. Duchnowski, S. Manke. Multimodal interfaces. *Artificial Intelligence Review*, 10, 299-319. 1996.

[Yang98] J. Yang, W. Lu, A. Waibel. Skin-color modeling and adaptation. *Proceedings Third Asian Conference on Computer Vision*, Vol. II, 142-147. 1998.