INTERFAZ MULTIMODAL BASADA EN GESTOS Y VOZ APLICADO A GRÚAS PÓRTICO

MULTIMODAL INTERFACE BASED ON GESTURES AND VOICE APPLIED TO PORTICO CRANES

Recibido: 06/07/07
Aceptado: 23/07/07

Alberto Isasi Andrieu
Ingeniero Industrial
Unidad Infotech **Tecnalia**

Artzai Picón Ruiz
Ingeniero Industrial
Unidad Infotech **Tecnalia**

Aránzazu Bereciartúa Pérez
Lic. en Ciencias Físicas
Unidad Infotech **Tecnalia**

José Ángel Gutiérrez Olabarría
Lic. en Informática
Unidad Infotech **Tecnalia**

RESUMEN

Hoy en día, los controles de los sistemas industriales siguen realizándose mediante teclados o mandos manuales, que normalmente requieren una formación específica para su manejo, además de no ser siempre intuitivos.

En este artículo se presenta un interfaz nuevo e intuitivo que mantiene la robustez, seguridad y eficiencia de los controles tradicionales aplicado en este caso al manejo de grúas puente.

Esta arquitectura emplea comandos de voz y gestos corporales estándar en este campo como medio de interactuar con el sistema objetivo.

Palabras clave: Multimodal, gestos, visión estéreo, comandos de voz.

ABSTRACT

Nowadays, industrial Control Systems are based on keypads and wired joysticks that usually are unfriendly and not intuitive, generally demanding practical experience. In this paper a new and more natural and intuitive interface is presented, keeping the present robustness and efficiency of traditional controls. The proposed interface is designed to control an overhead crane by using voice commands and gestural commands, since this language is more natural and intuitive for the users

***Key words**: Multimodal, gestures, stereo vision, voice commands.*

1.- INTRODUCCIÓN

La mayor parte de las máquinas y equipos industriales están controlados mediante teclados y mandos manuales que necesitan un uso activo por parte del operario. Además, es necesario que el operario esté en el entorno del interfaz interactuando mecánicamente con él.

En una instalación industrial hay, además, otras tareas como las logísticas de mercancías, que requieren la conducción por un operario o su manejo mediante controles remotos. En el mejor de los casos, estos dispositivos son inalámbricos, pero obligan a cargar con ellos y manejarlos con guantes y otros EPI. En el peor de los casos, la comunicación se lleva a cabo gritando a viva voz las instrucciones entre operarios teniendo en cuenta, además, que son entornos en los que el nivel de ruido es alto y la visibilidad no siempre está garantizada (vapores, zonas con diferente iluminación)

Todo esto implica que la comunicación en estos entornos sea dura y no siempre esté bien gestionada. Estos escenarios presentan distintas necesidades de optimización del trabajo (velocidad de respuesta) y seguridad para los usuarios (objetos peligrosos en las proximidades, alertas de seguridad que no son atendidas...)

Se hace, por tanto, necesario desarrollar aplicaciones que aporten interfaces cognitivos entre los hombres y las diferentes máquinas permitiendo una interacción que incluya el concepto de "inteligencia" y que facilite y haga más cómodas y seguras las tareas.

Los campos de aplicación de este tipo de tecnologías son enormes (puertos, aeropuertos, industria, turismo, seguridad...) y se hizo necesario

limitar los escenarios de aplicación para obtener un sistema robusto. Concretamente se seleccionó el ámbito industrial y, como ejemplo, se implementó el manejo de grúas puente.

2.- REQUISITOS DEL SISTEMA

El prototipo va a trabajar en un entorno industrial recibiendo información desde diferentes usuarios que se encuentren en la zona de trabajo. Por tanto, se deben cumplir los siguientes requerimientos básicos:

- El sistema tiene que entender los comandos gestuales, concretamente los realizados por medio del cuerpo (posición), manos y voz.
- El sistema trabajará bajo unas condiciones de iluminación muy variables y, en general, no serán las óptimas.
- Habrá que evitar posibles errores de entendimiento entre los comandos dados por el usuario y los interpretados por el sistema

Para resolver estas necesidades básicas, se tendrán en cuenta los siguientes aspectos:

- Se implementará el modelo de interacción estándar "UNE 003: *Código de señales para el manejo de grúas*" (UNE 58000:2003) para la interacción con grúas (**Isidro** (9)).

- El sistema contará con un par de cámaras infrarrojas y un foco infrarrojo para poder obtener medidas 3D con gran precisión; el usuario vestirá unos marcadores especiales sensibles a la luz infrarroja en sus guantes, casco de trabajo y buzo. De esta forma podrán obtenerse unos buenos modelos de gesto del usuario (**Cipolla** y **Pentland** (5).

- Como medida de control, la interpretación de la información recibida será realizada de forma redundante. El comando será dado tanto con un gesto como con una orden de voz. De esta forma, si aparecen comandos contradictorios el sistema contará con una información adicional (como un micrófono especial de laringe, ver *Nueva Electrónica* (7) and *NitSpy* (8)) para poder responder siempre en beneficio y seguridad de los usuarios.

3.- FUNCIONAMIENTO DEL SISTEMA

3.1.- Visión estereoscópica

Como ya se ha mencionado, el sistema contará con un par de cámaras infrarrojas (IR) para contar con una plataforma estereoscópica montada con un ángulo fijo para mantener constantes las relaciones geométricas. Además, el potencial del equipo no se ve reducido y las siguientes tareas, como la calibración y la búsqueda del usuario son más sencillas. Entre las dos cámaras se sitúa un foco IR que emitirá la cantidad de luz necesaria en la dirección de las cámaras para obtener el brillo óptimo de los marcadores.

En la figura 1 se presenta el diseño del prototipo. Este sistema está montado en el pilar del puente grúa de forma que podrá realizar un giro de casi los 360º sobre su propio eje para realizar la búsqueda de usuario en el inicio de la ejecución de comandos. La implementación de laboratorio no emplea esa posibilidad.

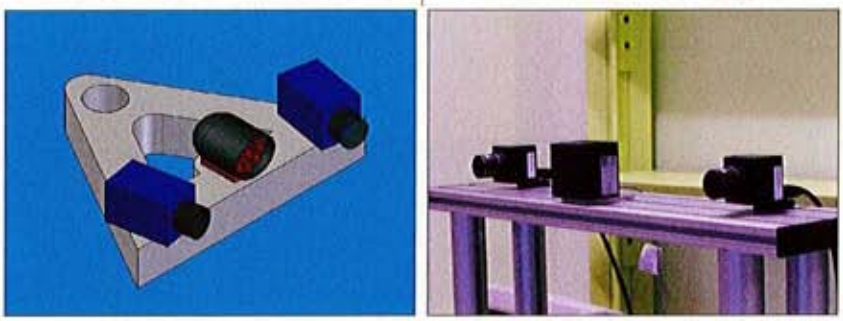

Figura 1: Conjunto de cámaras y foco.

La detección de los diferentes comandos gestuales dados por el usuario se producirá buscando los marcadores IR, situados en los guantes, casco y otros EPI, de forma que es más sencillo realizar las tareas de

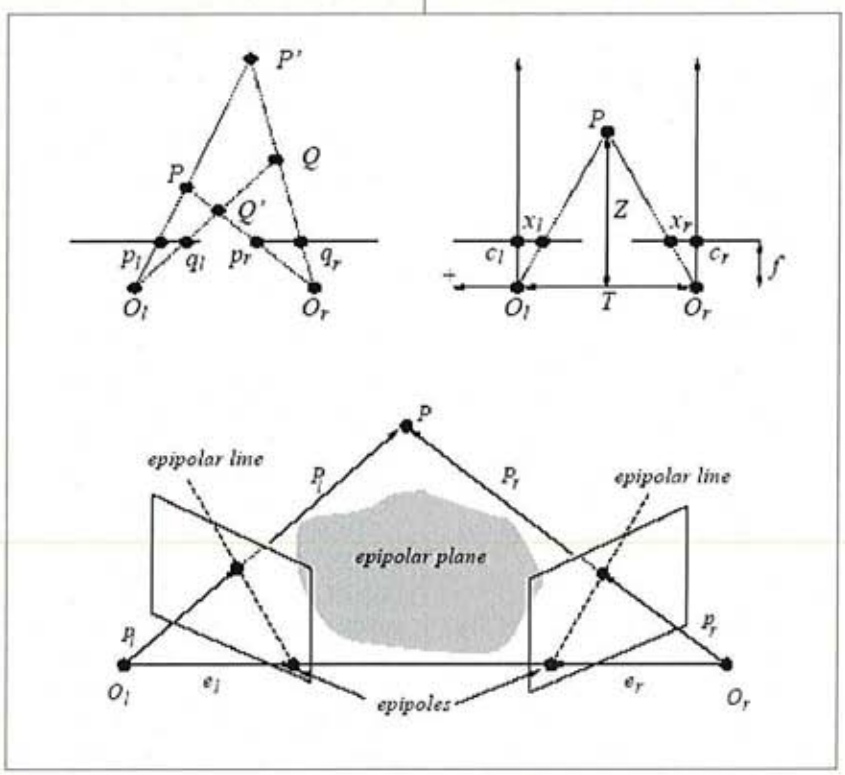

Figura 2: Fundamentos estéreo

segmentación ya que el fondo podrá ser fácilmente eliminado de la imagen y la robustez del sistema crecerá.(**Nikoladis** (4) , **Cipolla** y **Pentland** (5), **Fangeras** (6)).

Las imágenes adquiridas por las cámaras serán preprocesadas de forma que se pueda extraer la información de los marcadores IR del resto de la imagen (segmentación). Los puntos obtenidos en cada una de las imágenes (información 2D) se usarán para obtener las coordenadas en el espacio (información 3D) de la posición del usuario, utilizando las fórmulas de los sistemas de visión estéreo ((10). Teniendo en cuenta los puntos anteriores y los actuales, se puede aplicar una gramática gestual para determinar si se está produciendo uno de los comandos gestuales (**Sharma** *et al.* (1), **Corradini** y **Cohen** (2), **Lenman** *et al* (3). Si el sistema identifica un comando gestual, el proceso continúa. En caso contrario, el punto será almacenado en la lista de puntos anteriores para tenerlo en cuenta en el siguiente ciclo de procesado.

Siguiendo las ecuaciones de (11) se deben resolver dos problemas fundamentales en el procesamiento estéreo:

- Problema de correspondencia entre ambas imágenes.
- Problema de reconstrucción a partir de las disparidades.

Para ello, a partir de la figura 2, usando semejanza de triángulos, hay que resolver

$$\frac{T+x_l-x_r}{Z-f}=\frac{T}{Z} \quad \rightarrow \quad Z=\frac{fT}{x_r-x_l}$$

siendo necesario conocer d = Xr-X_l, f, T, C_l y Cr para calcular Z. Para ello es necesario resolver parámetros extrínsecos tales que

- Los puntos P_l y Pr se refieren al mismo punto en 3D respecto a las cámaras derecha e izquierda.
- Las relaciones entre ellos están establecidas por las matrices de traslación y rotación T y R: Pr = R (P_l-T).
- Los puntos de imagen pl y pr se relacionan con los puntos 3D por las ecuaciones de perspectiva p_l= fl x P_l / Z_l, pr= fr x Pr / Zr
- Restricción epipolar: cada punto 3D define un plano epipolar que intersecciona con la imagen a lo largo de la línea epipolar.
- Para el cálculo de la línea de línea epipolar es necesario el cálculo de la "matriz esencial" que define la relación entre el punto de imagen en coordenadas de cámara y la línea epipolar y la "matriz fundamental" que define la relación entre un punto de imagen en coordenadas pixel y la línea epipolar.
- Para ello se emplean métodos de calibración basados en dianas (Fig. 3).

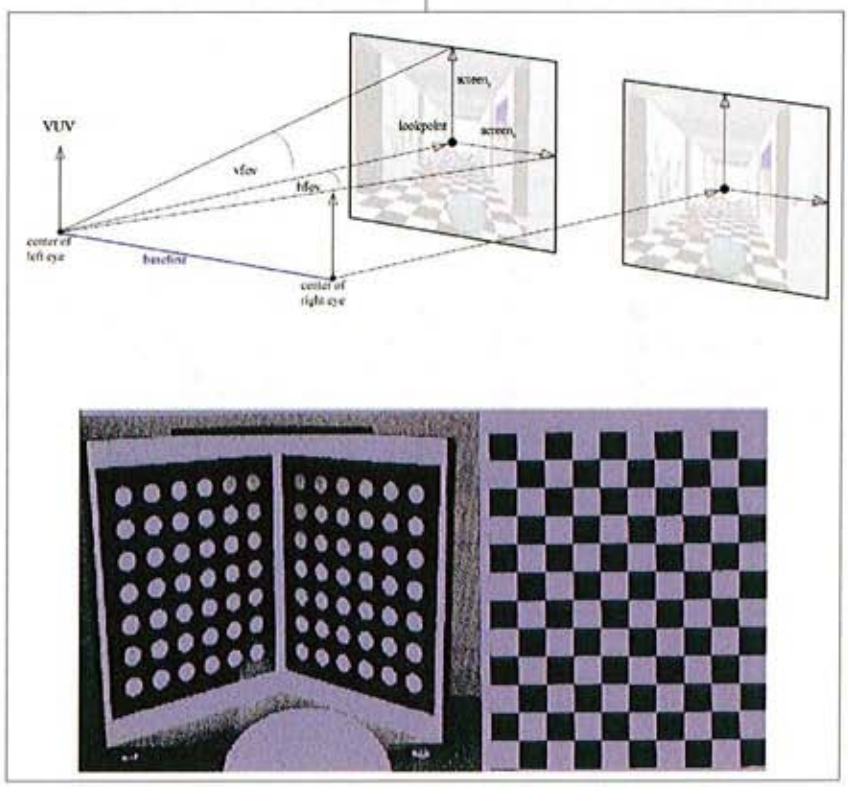

Figura 3: Principio de estéreo visión y diana

3.2.- Procesamiento de comandos de voz

Al mismo tiempo, y de forma paralela, se procesará la información recibida por los comandos de voz. Esta información será procesada usando las librerías de tratamiento de voz y la información procesada será introducida en una gramática de voz. Si el comando existe, el algoritmo continuará.

Los comandos de voz se darán con un micrófono de laringe (7) (8) por su mayor inmunidad al ruido ambiental que los micrófonos normales.

Los comandos gestuales y de voz estarán sincronizados con una señal de reloj para obtener un procesamiento paralelo entre ellos, de forma que las dos señales vayan sincronizadas en el procesamiento y no interfieran nuevas entradas de uno de los tipos de comando sobre el otro.

Como medida de seguridad, y para asegurar el entendimiento del comando, el sistema comparará el comando gestual con el comando de voz y, en caso de que exista alguna contradicción entre ellos, el sistema cambiará a un modo de error. En este modo, el sistema primará la seguridad del usuario y la respuesta podrá ser variable en función de cada situación.

Una vez que cada uno de los comandos (gesto y voz) ha sido obtenido, se comparan y se comprueba que los dos son iguales. Si, tanto la voz como el gesto, hacen referencia al mismo comando de control, la señal de control correspondiente será enviada al puente grúa, mientras que, si

son diferentes, el sistema no hará nada puesto que el sistema no tiene la seguridad necesaria para determinar cuál de los dos ha sido el adecuado.

Figura 4: Micro de laringe y marcadores para los EPI

En la figura 6 se muestran los diferentes estados del sistema.

En este diagrama se pueden ver cinco estados diferentes:

- Estado inicial: El sistema estará esperando por el comando de inicio. Después de recibir este comando, el sistema lanzará la tarea de localización de usuario.

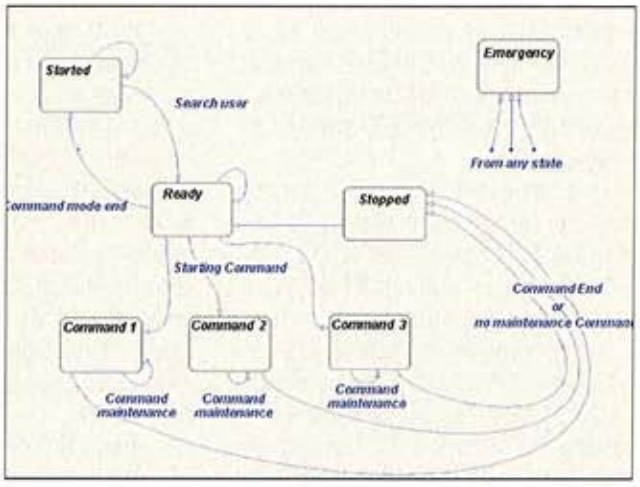

Figura 6: Diagrama de estados

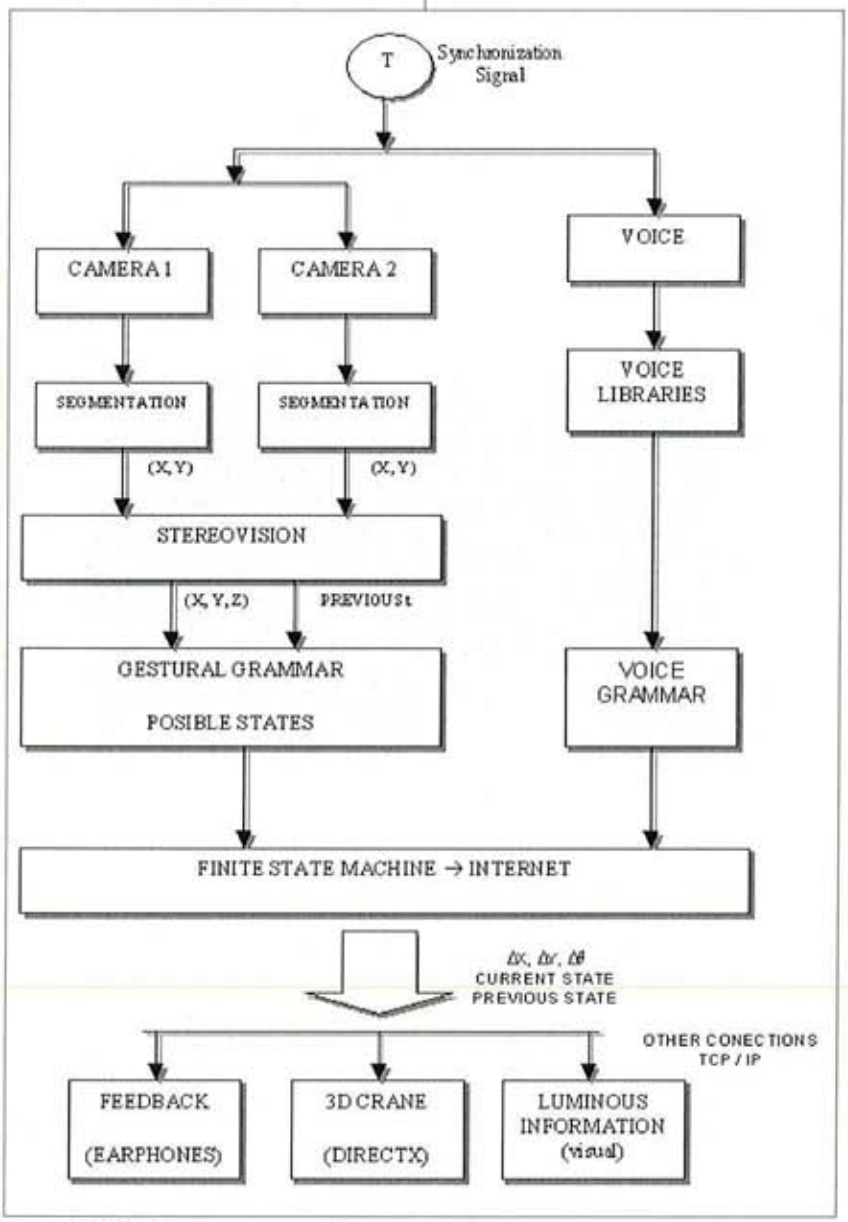

Figura 5: Arquitectura propuesta para el sistema

- Estado preparado: Una vez localizado el usuario, el sistema esperará a los comandos de control.

- Estado de mantenimiento de comandos: Una vez que el primer comando ha sido recibido, si éste pertenecía a la lista de comandos de movimiento, el sistema se mantendrá a la espera de nuevos comandos de movimiento en este estado o hasta que llegue el comando de parada.

- Estado de parada: Una vez recibido el comando de detención, el sistema detendrá las órdenes de movimiento y cambiará el estado a "Estado preparado".

- Estado de emergencia: El sistema puede llegar a este estado desde cualquiera de los otros estados y no es necesario que se produzca la redundancia de comandos a la hora de aceptar el comando de parada de emergencia. Con esto, en caso de un peligro inminente, la acción del usuario puede ser más natural y no forzada. Para ello, se habilitarán más de un comando de voz para realizar esta parada de emergencia.

En la Tabla 1 se muestran las instrucciones básicas para controlar el movimiento del puente grúa.

La implementación de este primer prototipo será llevada a cabo de forma virtual. La razón de ello reside en que, en esta fase, el objetivo es diseñar una interfaz robusta y amigable

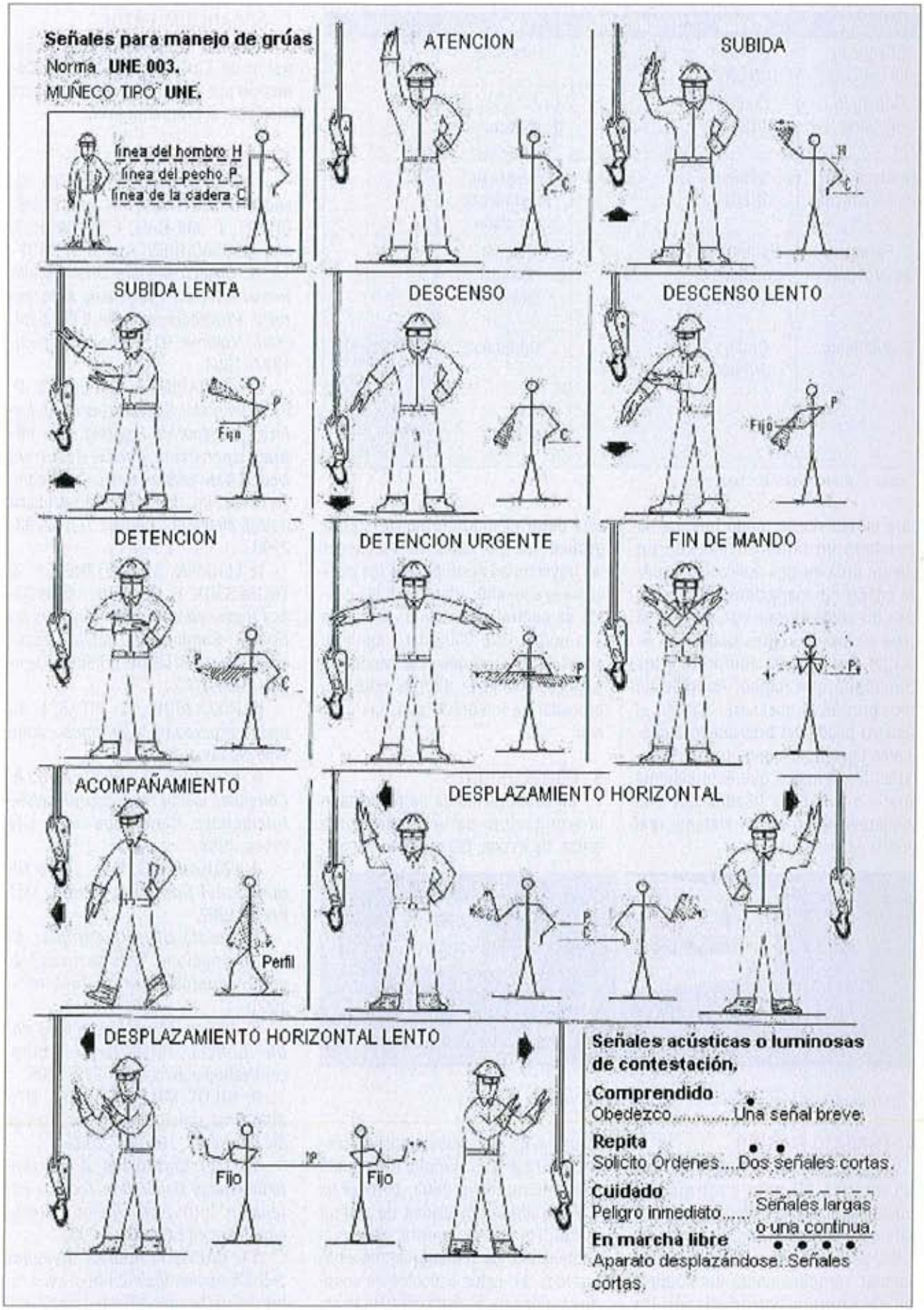

Figura 7. UNE 58000:2003 Comandos para la interacción con una grúa

Comando	Gesto	Redundancia	Voz	Notas
Búsqueda de usuario	Atención (UNE003)	⊙	"Búscame"	
Comando de inicio	Código UNE003	⊙	"Grúa" + Palabra de movimiento asociada	
Comando de mantenimiento	Código UNE003	⊙	Palabra de movimiento asociada	
Parada de emergencia	Parada de emergencia (UNE003)	+	"Detente", "Peligro" "Ayúdame" ...	
Seguimiento	Código UNE003	⊙	"Continúa"	Este comando debe comenzar con el gancho de la grúa cerca del usuario para evitar posibles obstáculos

Tabla 1. Instrucciones básicas

para el usuario de modo que, se ha diseñado un taller virtual con un puente grúa en una aplicación capaz de buscar los marcadores IR y escuchar los comandos de voz, y mover la grúa en los diferentes grados de libertad que soporta conforme a los comandos que recibe. Además de ello, gracias al escenario virtual, el sistema puede ser presentado a diferentes tipos de usuarios para presentarles las ventajas que este sistema aporta y evitar los riesgos que una implementación en un sistema real podría generar.

Figura 8: Taller virtual del puente grúa y entorno de pruebas

4.-TRABAJO FUTURO

Actualmente se ha implementado un interfaz 3D para controlar un puente grúa virtual usando gestos estáticos similares a los de la norma UNE, pero el objetivo final es implementar completamente los códigos de esta norma, introduciendo los gestos en movimiento. Para ello, se está desarrollando un nuevo interfaz gestual que sea capaz de reconocer las trayectorias descritas por los marcadores e identificarlas como los gestos de control del usuario, conforme a la norma UNE 0003 del control de grúas. Y, una vez que esta interfaz se haya probado en el entorno virtual, se procederá a integrarlo con una grúa real.

5.-CONCLUSIONES

En este artículo se ha presentado la arquitectura de un control para grúas, de manos libres. Este sistema, actualmente sólo trabaja en un escenario virtual sin cumplir totalmente con la norma UNE 0003, pero se ha demostrado la posibilidad de realizar el control de un puente grúa real cumpliendo los requerimientos establecidos. En estos aspectos se continúa la línea de investigación en la actualidad.

6.-AGRADECIMIENTOS

Agradecimientos al **Ministerio Español de Turismo, Industria y Comercio** por el soporte a este proyecto mediante el programa *Profit.*

7.-BIBLIOGRAFÍA

1. SHARMA, R. and YEASIN, M. and KRAHNSTOEVER, N. and RAUSCHERT, I. and CAG, I, BREWER, I. and MACEACHREN, A. and SENGUPTA, K. *Speech-Gesture Driven Multimodal Interfaces for Crisis Management.* Proceedings of the IEEE, Sept. 2003, Volume: 91, Issue 9, page(s): 1327- 1354.

2. CORRADINI, A. and COHEN, P. R. *Multimodal Speech-Gesture Interface for Handfree Painting on a Virtual Paper using Partial Recurrent Neural Networks as Gesture Recognizer.* Proc. Int. Joint Conf. on Artificial Neural Networks (IJCNN '02), 2293-2298.

3. LENMAN, S., BRETZNER, L. & THURESSON, B. *Computer Vision Based Recognition of Hand Gestures for Human -Computer Interaction.* Technical report TRITANA -D0209, CID-report, June 2002

4. NIKOLAIDIS, N., PITAS, I. *3D image processing algorithms* , John Wiley & Sons, 2001.

5. CIPOLLA, R., PENTLAND, A. *Computer Vision for Human-Machine Interactions.* Cambridge University Press. 1998.

6. FAUGUERAS, O.D. *Three Dimensional Computer Vision.* MIT Press. 1992.

7. Revista *Nueva Electrónica. Nº* 89). Laringófono. http://www.ea8zq.com/esquemas2/laringofono.htm. 1991

8. *NitSpy. Micro-laringófono doble policial.* http://www.nitspy.com/eshop.asp?cod=MAE19. 2006

9. SILOS MILLAN, Isidro. *NTP 208: Grúa móvil.* http://www.mtas.es/insht/ntp/ntp_701.htm. 2003

10. MEI, Christopher. *A new Omnidirectional Calibration Toolbox Extension.* http://www.vision.caltech.edu/bouguetj/calib_doc/. 2005

11. CALWAY, Andrew, *Dynamic 3-D Computer Vision.* http://www.cs.bris.ac.uk/Teaching/Resources/COMS30121/slides3D/ ■