REVISTA COLOMBIANA DE COMPUTACIÓN
Volumen 14, número 2
Págs. 141 - 161

Propuesta de Interacción Multimodal: Caso de estudio Navegación en Mapas

Fernando De la Rosa[1]‡, Carlos Duque[1], José T. Hernández[1].

Fecha de recibido: 01/11/2013 Fecha de Aprobación: 09/12/2013

Resumen

Este trabajo presenta una propuesta del componente de interacción de usuario multimodal para ser integrado principalmente con aplicaciones de computación visual 2D/3D. El componente considera múltiples canales de entrada de información funcionando simultáneamente. Sobre cada canal de entrada se define un proceso de tratamiento de información que corresponde a una modalidad de interacción disponible al usuario. La información resultante de las diferentes modalidades se integra para definir acciones/comandos conjuntos de interacción hacia la aplicación computacional. La interacción multimodal busca ser más natural para el usuario, lo cual significa una interacción más fácil y agradable. La propuesta fue validada integrando una componente de interacción multimodal basada en gestos manuales y voz con la aplicación Bing-Maps para realizar la tarea de navegación en mapas 2D en un ambiente de visualización de mediano formato. El prototipo de integración fue validado por un conjunto de usuarios y se obtuvieron resultados preliminares de su funcionalidad y de la experiencia de usuario resultante.

Palabras Clave: *Interacción Hombre-Máquina, interacción multimodal, interacción gestual, interacción por voz, navegación en mapas.*

Abstract

This work presents a proposal of a multimodal user interaction component to be integrated mainly with 2D/3D visual computing applications. The component considers multiple input channels of information working simultaneously. On each input channel is defined an information process corresponding to an interaction modality available to the user. The information resulting from the

1 Grupo de Investigación IMAGINE, Departamento de Ingeniería de Sistemas y Computación, Universidad de los Andes, Bogotá – Colombia, {fde, jhernand}@uniandes.edu.co, cduquer@gmail.com.

different modalities is integrated to define joint interactive actions/commands for the computer application. The multimodal interaction aims to be more natural for the user, which means an easier and pleasant interaction. The proposal was validated by integrating the interactive multimodal component based on hand gestures and voice with the Bing-Maps application to accomplish navigation in 2D maps inside a medium size visual environment. The integrated prototype was validated with a set of users and initial results were obtained regarding its functionality and the user experience.

Keywords: *Human computer interaction, multimodal interaction, gestual interaction, voice interaction, navigation in maps.*

1. Introducción

La interacción humana con el medio circundante y con otros seres humanos es sin duda multimodal, es decir que se usan múltiples canales de comunicación para transmitir información (*e.g.*, los gestos manuales, la voz, la escritura, la mirada) o para adquirir información (a través de la vista, el oído y el tacto). En el contexto de Interacción Hombre-Máquina (IHM o *Human Computer Interaction* HCI), Oviatt [1] hace referencia al concepto de interacción multimodal cuando el usuario dispone de múltiples modos (simultáneos) (*e.g.* gestos {manuales, faciales, cabeza, corporales}, voz, escritura, tacto y/o mirada) para interactuar de manera coordinada con un sistema/aplicación computacional multimedia. En particular, el desarrollo y uso de aplicaciones computacionales con contenido visual 2D/3D (*e.g.* videojuegos, uso y consulta de mapas, simuladores para tratamientos médicos, ambientes de entrenamiento, visualización de grandes volúmenes de información) favorecen la aplicación de interacción multimodal o basada en múltiples canales de información. El acercar la interacción de los usuarios de una aplicación computacional a mecanismos y metáforas más naturales favorece la usabilidad y la satisfacción para los usuarios. Los recientes avances tecnológicos en las áreas de diseño de nuevos dispositivos de interacción y visualización, procesamiento de señales, visión por computador e interacción Hombre-Máquina han hecho posible el estudio y desarrollo de interfaces multimodales.

Este artículo es una extensión de la arquitectura propuesta en [2], en el cual se hace un análisis más completo de los trabajos relacionados, se detalla cada uno de sus componentes y se profundiza en el protocolo de pruebas ejecutado. En este artículo se propone una componente de interfaz de usuario multimodal, la cual permite la integración de

múltiples modalidades, cada modalidad con su procesamiento especializado, para luego integrar sus resultados y definir acciones/comandos conjuntos que se comunican a la aplicación. La expresión de la interacción de la aplicación se define por una máquina de estados finitos donde se identifican los posibles estados de la aplicación y las transiciones entre estados corresponden a las acciones/comandos que debe reconocer la capa de interacción a partir de las entradas dadas por el usuario.

En aplicaciones computacionales interactivas donde la visualización se basa en escenarios y/u objetos 2D/3D, las acciones que puede realizar el usuario se pueden agrupar bajo tres operaciones básicas: la selección y la manipulación de objetos al interior de los escenarios y la navegación/exploración del escenario. Para la validación de la propuesta de este trabajo se desarrolló una componente de interacción multimodal basada en un conjunto de gestos manuales y comandos de voz para realizar las acciones comunes en la navegación de mapas 2D. Esta nueva componente se integró con la aplicación Bing-Maps [3]. Egenhofer [4] y McGee *et al.* [5] sugieren que las interfaces multimodales son más eficientes para interactuar con información geo-espacial que la interacción unimodal. La interacción por voz provee una forma directa y efectiva de expresar acciones y sustantivos, pero falla cuando se tienen que definir posiciones. Por esto, los gestos pueden proveer una modalidad alterna/complementaria efectiva que es más apropiada para expresar relaciones espaciales y es menos propensa a errores que si se expresara sólo en palabras.

Este artículo está organizado de la siguiente manera: En la sección 2 se describe el problema de estudio. Posteriormente se presenta el estado del arte de aplicaciones basadas en interacción multimodal (sección 3) seguido de la descripción detallada de la componente multimodal propuesta (sección 4). El caso de estudio de validación que detalla la forma de interacción multimodal gestos y voz integrada a la aplicación Bing-Maps se explica en la sección 5. Las pruebas de usuario y los resultados preliminares de la evaluación con usuarios se presentan en la sección 6. Por último se presentan las conclusiones y el trabajo futuro.

2. Problema de Estudio

Las interfaces de interacción tradicionales para aplicaciones de computación visual 2D/3D (*e.g.*, navegación en mapas) limitan la usabilidad al no ser lo suficientemente intuitivas a los usuarios y/o al ralentizar la interacción como resultado de intentar encajar los grados de

libertad (g.d.l.) de los datos espaciales en 2D (2 g.d.l. en posición + 1 g.d.l. en rotación) o de los datos espaciales en 3D (3 g.d.l. en posición + 3 g.d.l. en rotación) en los 3 g.d.l. (2 g.d.l. de posición + 1 g.d.l. de rotación) que se logra en una proyección 2D de una escena usando un dispositivo apuntador (tipo ratón convencional) controlado por una mano del usuario. Sin embargo, la voz y el sistema motriz humanos permiten naturalmente describir acciones a través del habla o usando las manos en tres dimensiones por lo que el paradigma del ratón no está aprovechando completamente la capacidad humana para generar información.

Una interfaz de usuario multimodal que use habla y gestos de manos como la que se presenta en este trabajo, ofrece ventajas sobre los sistemas que usan una sola modalidad o usan dispositivos estándares (ratón y teclado).

3. Trabajos Relacionados

WallBoard [6] es un dispositivo basado en la metáfora del tablero de oficina, con herramientas físicas (*e.g.* marcadores y borradores) y virtuales (*e.g.* lentes y dispositivos de medición) que tienen capacidades de uso similares a las herramientas asociadas a un tablero común. El *Wallboard* presenta un caso de manipulación de información espacial con una interfaz de usuario multimodal a partir de las siguientes modalidades: gestos con las manos, gestos de contacto, movimiento de cabeza-ojos y comandos de voz. El sistema reconoce la combinación de diferentes modalidades de interacción. El *WallBoard* se propone para el caso de aplicaciones SIG (Sistemas de Información Geográfica), como un ambiente multi-usuario colaborativo donde se pueden realizar operaciones de selección, zoom (manejo del nivel de detalle), desplazamiento, rotación y navegación.

MATCH -Acceso multimodal a la ayuda de una ciudad- [7] es una guía interactiva de la ciudad y aplicación de navegación que provee información de restaurantes y el metro subterráneo de New York. MATCH permite realizar las operaciones de selección de restaurantes en un mapa 2D, obtener información sobre dichos restaurantes, navegar en un mapa 2D y mostrar rutas hacia las estaciones de metro subterránea. *MATCH* posee una interfaz gráfica dinámica que tiene efectos sobre los mapas desplegados en una tableta digital. Como entradas, el sistema reconoce comandos de voz, gestos de escritura con lápiz y/o combinación de ambos y como salidas usa síntesis de voz sincronizada con imágenes dinámicas. La representación de la multimodalidad de una aplicación interactiva se realiza utilizando gramáticas libres de contexto y métodos de estados finitos.

Larsen *et al.* [8] presentan un entrenador automático para el aprendizaje del juego de billar (APT). El entrenador automático puede presentar diferentes ejercicios al usuario. La información de un ejercicio y las diferentes acciones asociadas al ejercicio se proyectan sobre la mesa de juego. El usuario puede interactuar con la información proyectada (usando gestos con una mano) para seleccionar diferentes funcionalidades del entrenador. El usuario también puede comunicarse a través de comandos de voz. El entrenador automático puede reconocer el resultado de una jugada del usuario a través del reconocimiento de las bolas de billar en la mesa aplicando técnicas de procesamiento de imágenes. El entrenador automático también usa voz sintética y sonidos informativos para dar instrucciones al usuario. El sistema automático de entrenamiento ha sido probado por múltiples usuarios con un gran nivel de aceptación.

Ruiz *et al.* [9] presentan una plataforma que integra aplicaciones 2D/3D con interacción basada en gestos. La interacción gestual se basa en un proceso comprendido como el flujo de la información proveniente de los dispositivos de interacción que utiliza el usuario, su interpretación gestual y su transformación a una acción interpretable por la aplicación 2D/3D. La plataforma propuesta se basa en la definición de tres etapas distinguibles para la integración de la interacción gestual con una aplicación visual interactiva: gestualidad (adquisición de datos e identificación de gestos), interpretación (transformación de los gestos en valores numéricos y normalizados) y procesamiento (transformación hacía las acciones y sus parámetros que define la aplicación computacional y que deben ser comunicados a la componente aplicación para su actualización). Una posibilidad que ofrece la plataforma es que la etapa de gestualidad puede implementarse bajo diferentes contextos físicos (insfraestructura de interacción) lo que permite integrar una misma aplicación 2D/3D bajo diferentes modalidades de interacción (*e.g.* gestos manuales, gestos de contacto, gestos con *joystick*). La plataforma se evaluó con una aplicación de navegación de mapas 3D de la ciudad de Bogotá donde se definen 3 g.d.l. de desplazamiento, 2 g.d.l. de rotación y 1 cambio de estado temporal.

Ying [10] propone una plataforma de interacción multimodal basada en modelos de Markov ocultos que permitan un análisis continuo de gestos para lograr una interacción natural con el usuario. Bajo la plataforma, los gestos y posturas de la mano junto con la voz son la base de la interacción. El trabajo propone realizar un análisis para reconocer movimientos de la mano (no intencionales), gestos manipulativos y gestos comunicativos. También se considera la voz como un canal de

comunicación disponible al usuario. En el aspecto multi-modal se busca distinguir cada modalidad de interacción y lograr determinar los momentos en que las modalidades se consideran alineadas para complementarse. El trabajo realiza la propuesta de la plataforma pero no se presenta un prototipo disponible para realizar pruebas con usuarios.

El trabajo de Daiber *et al*. [11] describe una propuesta de interacción multi-modal basada en gestos y posturas de ambas manos. Las posturas de las manos son útiles para describir acciones del usuario largas en el tiempo. La aplicación de prueba está relacionada con la selección, exploración y navegación en un sistema de información geográfica 2D. La información al usuario se muestra sobre una mesa de proyección, la cual puede centrarse en un área de interés y ser desplegada en un dispositivo móvil. En la nueva visualización se puede completar el nivel de detalle de la información. Los gestos y posturas de las manos se reconocen a través de una cámara. Se reportan pruebas preliminares del uso de un prototipo por parte de un conjunto de usuarios.

De nuestra parte, el trabajo que presentamos en este artículo propone una componente de interacción de usuario multimodal que se integra a una aplicación computacional con contenido visual 2D/3D. En el caso de estudio considerado, esta componente se desarrolló para reconocer gestos/movimientos del usuario con cada mano por separado y reconocer un subconjunto de comandos de voz. La componente resultante se integró a la versión actual de Bing-Maps para navegar en mapas 2D usando un dispositivo Kinect (Microsoft) y un ambiente de visualización de mediano formato. El prototipo desarrollado fue evaluado por un conjunto de usuarios en criterios como la "correspondencia" de los gestos y comandos de voz con los comandos y/o acciones en la aplicación computacional, el tiempo de respuesta (latencia) de la aplicación, la facilidad de uso del prototipo y la satisfacción de realizar la tarea de navegación en un mapa usando interacción multimodal. Las características de nuestra propuesta de interacción multimodal y los resultados obtenidos se presentan en las secciones a continuación.

La Tabla 1 presenta un comparativo entre los trabajos relacionados, incluyendo la propuesta de interacción de usuario multimodal presentada en este artículo. Las características de comparación son:

Referencia	Concepto/ Prototipo y Aplicación	Operaciones Básicas	Dispositivos Entrada	Modalidades Interacción	Dispositivos Salida
[6]	Concepto *Wallboard:* sistema información geográfica 2D	Selección, Navegación	tablero digital, marcador, borrador, movimiento cabeza-ojos	Gestos por contacto + Gestos manos + Gestos cabeza-ojos +Voz	Tablero digital
	Prototipo	Selección,	Tableta digital,		Tableta digital
[7]	MATCH: Información sitios ciudad 2D	Navegación	micrófono	Gestos escritura + Voz	+Voz sintética
[8]	Prototipo Entrenador Billar 2D	Selección	cámara, micrófono	Gestos manos + Voz	Proyector + Voz sintética + sonidos informativos
[9]	Prototipo Mapas 3D	Navegación 3D	OptiTrack (cámaras infrarojas)	Gestos manos	Mesa proyección
[10]	Concepto Juego	Manipulación	Kinect (incluye micrófono)	Gestos manos + Voz	Proyector
[11]	Prototipo *TouchPosing:* sistema información geográfica 2D	Selección, Manipulación, Navegación	Cámara infraroja, Mesa sensible contacto	Gestos manos + Tacto	Mesa proyección + tableta digital móvil
Propuesta Artículo	Prototipo Mapas 2D	Navegación	Kinect (incluye micrófono)	Gestos manos + Voz	Proyector /LCD

Tabla 1. Comparativo de características de interacción de usuario en los trabajos relacionados.

4. Arquitectura Propuesta

La arquitectura de Interacción Multimodal propuesta está compuesta por tres componentes (Figura 1): Dispositivo, Reconocimiento y Aplicación. La arquitectura permite el desarrollo de interfaces de interacción unimodal; las modalidades de entradas son reconocidas en paralelo y procesadas mediante el desarrollo de componentes especializados. Los resultados obtenidos de las interfaces unimodales implican representaciones parciales de la intención del usuario que son fusionados por el Integrador Multimodal con el fin de realizar una interpretación integrada multimodal que se transforma en comandos/acciones conjuntos de control que son enviados a la aplicación.

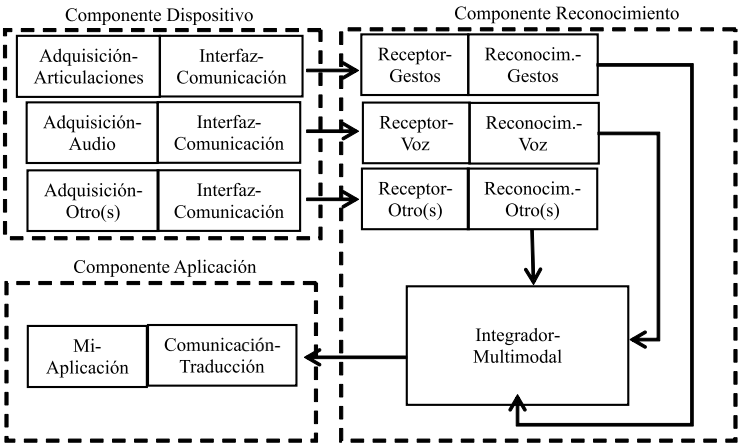

Fig. 1. Arquitectura de interacción multimodal basada en voz y gestos manuales.

4.1. Componente Dispositivo

El componente *Dispositivo* tiene como responsabilidad realizar la conexión con los dispositivos de interacción y realizar la adquisición de los datos brutos (*raw data*) que entregan estos dispositivos. Los datos recibidos por cada dispositivo deben ser transformados en información interpretable por el componente que reconoce las acciones que comunica el usuario.

En la arquitectura de interacción multimodal (Figura 1) este componente se especializa en dos módulos. El módulo *Adquisición-Articulaciones* que realiza la adquisición del dispositivo de seguimiento de movimiento (*tracking*) de los brazos. A partir de la información de movimiento se hace la transformación en el conjunto de valores/ángulos de las articulaciones superiores en el espacio 3D (sistema de referencia del usuario) y se genera un evento de actualización. Adicionalmente se tiene el módulo *Adquisición-Audio* que se conecta al dispositivo de audio y al detectar un comando de voz reconocido se genera un evento de actualización. Los módulos de Adquisición de información deben funcionar de forma independiente, por lo que requieren hilos de ejecución paralelos.

La información resultante de los módulos de Adquisición es enviada a la componente Reconocimiento a través del módulo *Interfaz-Comunicación*.

El componente *Dispositivo* puede ser extendido definiendo módulos de Adquisición para diferentes tipos de dispositivos físicos.

4.2.. Componente Reconocimiento

Este componente se encarga de recibir de forma independiente la información obtenida del *Componente Dispositivo*. Esta información es tratada de forma independiente para reconocer/identificar las acciones que comunica el usuario a través de cada dispositivo. El reconocimiento de una acción genera un evento que se envía al módulo *Integrador-Multimodal* con la información de esta acción.

En la arquitectura de interacción multimodal (Figura 1) este componente contiene dos módulos especializados: *Reconocimiento-Gestos* y *Reconocimiento-Voz*. El módulo *Reconocimiento-Gestos* define el protocolo para la detección de un gesto a partir de la información de los valores/ángulos de movimiento de los brazos. Este módulo se especializa en el reconocimiento de los diferentes gestos identificables por la arquitectura de interacción. En particular se desarrolló el módulo *Reconocimiento-Gesto-Swipe* para la detección del gesto movimiento de una mano (*swipe)* en las cuatro direcciones principales: abajo, arriba, izquierda y derecha (Figura 2). Procesa las posiciones de las articulaciones obtenidas del componente *Dispositivo* y decide si el usuario realizó un gesto *swipe* y su dirección. También se desarrolló el módulo *Reconocimiento-Gesto-FijarMira* para el reconocimiento de un punto que define el usuario con una mano (Figura 3).

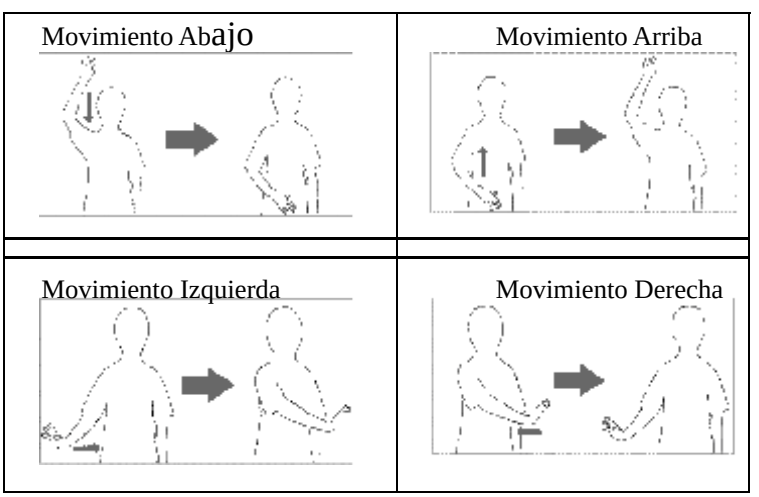

Fig. 2. Posibilidades del gesto de movimieto (*swipe*) para describir una dirección de Traslación y/o de Rotación.

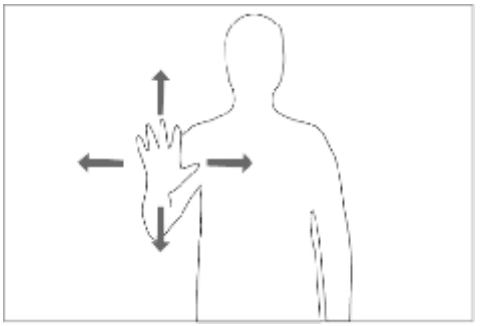

Fig. 3. Posibilidades del gesto "FijarMira" para controlar la mira utilizada para describir el punto de interés durante la operación de Zoom.

Para realizar el reconocimiento de gestos se hace uso de los datos de las articulaciones generadas por el *SDK* de Kinect para Windows [12] y de la detección algorítmica que consiste en definir reglas y condiciones que deben satisfacerse para identificar un gesto. Se deben capturar las *n* últimas posiciones de las articulaciones y se comparan con las condiciones y reglas definidas. La naturaleza simple de los algoritmos de detección limita el tipo de gestos detectables. La técnica algorítmica es apropiada para detectar movimientos lineales como el deslizar la mano de derecha a izquierda o de arriba hacia abajo (*swipe*) pero no para detectar el movimiento de un lanzamiento porque los movimientos del gesto de movimiento unidireccional del brazo son más simples y uniformes mientras que el movimiento de lanzamiento es más variable. Cada gesto a reconocer requiere de un algoritmo de detección ajustado al gesto, por lo tanto, hay un compromiso (relación inversa) entre el número de gestos a detectar y el desempeño/calidad de la detección de los gestos. Para favorecer el tratamiento homogéneo de un gesto se definen las siguientes funcionalidades comunes como:

- Capturar la posición de una articulación.
- Proveer un evento para señalar la detección de un gesto.
- Proveer un mecanismo para prevenir detectar gestos superpuestos, por ejemplo, definiendo un retraso temporal mínimo entre dos gestos.

También se tienen en cuenta las siguientes condiciones para la detección de un gesto de movimiento unidireccional con una mano (*swipe, FijarMira*)[13], comprobando que:

- La secuencia de puntos detectados está en progresión en la dirección requerida.

- La secuencia de puntos no está muy lejos de los otros en los demás ejes.
- El primer y el último punto describen una distancia suficiente para definir una dirección principal.
- El primer y el último punto fueron creados dentro de un lapso de tiempo delimitado para el gesto.

El módulo *Reconocimiento-Voz* hace uso de la plataforma de voz de Microsoft [14]. El reconocimiento de voz permite predefinir los comandos/*tokens* a detectar.

El módulo *Integrador-Multimodal* recibe los eventos de los reconocedores (gestos, comandos de voz y otras entradas) y los integra de acuerdo al momento en que se generaron creando la estructura (acción, parámetros) necesaria para realizar una acción conjunta. El *Integrador-Multimodal* usa las marcas de tiempo de los eventos de detección para determinar si un evento es parte de un comando/acción multimodal. Un comando/acción multimodal define una acción compuesta por dos o más entradas "simultáneas" dadas por el usuario (*e.g.* ("Mover", <dirección desplazamiento>) o (<dirección desplazamiento>, "Mover") donde "Mover" es un comando de voz y <dirección desplazamiento> es una dirección definida por el gesto de *swipe*). Para el diseño de las restricciones temporales el estado del arte actual de sistemas multimodales usa umbrales temporales fijos, que están basados en los patrones de integración observados en los usuarios. Oviatt *et al.* [15] muestran que el 88% de las veces el retraso entre el final de la señal de escritura del lápiz y el comienzo de la señal de voz es de 0 a 3 segundos, por lo que se usará ese intervalo para nuestro caso de estudio.

4.3. Componente Aplicación

Este componente es responsable de recibir los comandos/acciones (unimodales o multimodales) del *Integrador-Multimodal* (*Componente Reconocimiento*) a través de la Interfaz de Comunicación y los transforma en comandos/acciones y parámetros que la aplicación computacional pueda entender.

Los módulos principales que describen esta componente son *Comunicación-Traducción y Mi-Aplicación*. El módulo *Comunicación-Traducción* recibe los eventos del *Integrador-Multimodal* y los traduce a acciones y parámetros que la aplicación interactiva reconoce. Para realizar esta traducción, este módulo requiere tener conocimiento de los estados posibles de la aplicación, la(s) modalidad(es) de interacción reconocida(s) en cada estado para cada tarea/acción posible y los datos de entrada y sus rangos válidos asociados a cada acción/tarea. Esta

información define las posibilidades que ofrece una aplicación interactiva al usuario usando las modalidades de interacción disponibles. Esta información se representa a partir de una máquina de estados finitos, cada estado con las transiciones posibles a ejecutar desde dicho estado. El módulo *Mi-Aplicación* corresponde a la aplicación interactiva que ejecuta los comandos/acciones del usuario (con los parámetros necesarios) en el estado actual. Como respuesta, la aplicación ejecuta la tarea/acción respectiva en el estado actual, si requiere actualiza su estado y genera la realimentación hacia el usuario que puede incluir texto, audio, gráficas, imágenes y animación. Para el caso de estudio se desarrolló el módulo *Mi-Aplicación-BingMaps* como una especialización del módulo *Mi-Aplicación*, el cual ejecuta la aplicación Bing-Maps, con una interfaz gráfica modificada y que procesa la información resultante del módulo *Comunicación-Traducción*.

5. Caso de Estudio

La interacción de ratón/teclado de Bing-Maps define las siguientes acciones de navegación (Figura 4): el control de zoom permite acercarse (presionando el botón +) o alejarse (presionando el botón –) del centro de la visualización. El control de movimiento permite desplazarse sobre el mismo plano manteniendo presionadas las flechas de acuerdo a la dirección en que se quiere mover.

Fig. 4. Interfaz gráfica tradicional Bing-Maps (controles de navegación en la esquina superior derecha).

Como caso de estudio se propone una interfaz de interacción multimodal para la navegación de mapas en un ambiente de visualización de mediano formato usando Bing-Maps a partir de un conjunto de gestos y comandos de voz que incluye las siguientes acciones (Figura 5): zoom (acercar/alejar), movimiento en las cuatro direcciones principales (derecha/izquierda, arriba/abajo) y rotación con respecto a un eje. Para la navegación en mapas se tiene en cuenta que la

interacción con voz facilita la expresión de acciones y sustantivos y la interacción gestual facilita la expresión de relaciones espaciales. La figura 6 muestra un usuario interactuando con Bing-Maps a través de la componente de interacción multimodal.

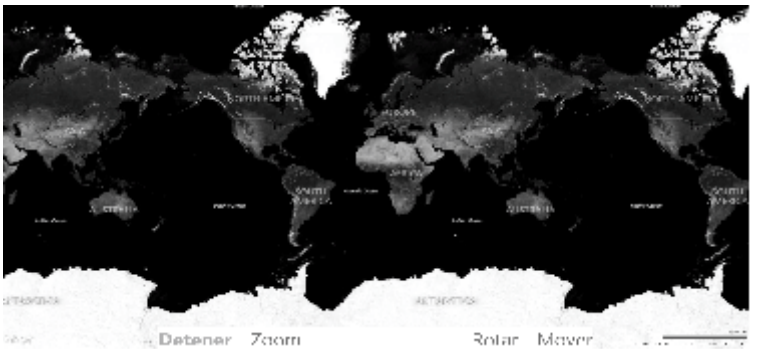

Fig. 5. Interfaz gráfica de la versión multimodal Bing-Maps que incluye el estado actual (zona inferior) resultado de la detección de un gesto y/o comando de voz.

Fig. 6. Usuario interactuando con la aplicación Bing-Maps a través de la nueva componente de interacción multimodal. El componente Dispositivo hace uso del dispositivo Kinect (Microsoft) para la adquisición de audio y del movimiento de brazos/manos. La visualización se realiza en un monitor LCD 42".

Chang y Bourguet [16] describen la utilidad de las máquinas de estados finitos como una herramienta de modelado para el diseño multimodal, donde se puede describir y entender fácilmente la estructura de las interacciones multimodales usando los conceptos de estado y acciones que provocan cambios de estados. Para el caso de estudio se diseñó la máquina de estados que modela la interacción multimodal a partir de las siguientes acciones (Figura 7):

Zoom: El comando de voz "Zoom" permite pasar al estado "EnEsperaDeAcción" lo que habilitará la mira en la visualización para señalar el punto de interés usando el gesto "FijarMira" (Figura 3). Al enunciar el comando de voz "Acercar" o "Alejar", la visualización se acercará o alejará con respecto al punto señalado hasta que el usuario enuncie el comando "Detener". Durante la acción "Acercar" o "Alejar" se puede alternar al estado opuesto enunciando el comando correspondiente. Para esta acción se necesita la estructura ("Acercar" o "Alejar", <punto de interés>) donde la acción se obtiene de un comando de voz y el <punto de interés> se adquiere del punto señalado con una mano por el usuario.

Movimiento (Traslación): Se debe enunciar el comando de voz "Mover" para llegar al estado "EnEsperaDeDirecciónDeMovimiento" y realizar el gesto de movimiento (*swipe)* en la dirección en que se quiera realizar el desplazamiento (Figura 2). Esto generará que la visualización se desplace de acuerdo a la dirección señalada hasta que el usuario enuncie el comando "Detener". Es posible cambiar la dirección del movimiento haciendo el gesto *swipe* en la dirección deseada. Para esta acción se necesita la estructura ("Mover", <dirección desplazamiento>) o (<dirección desplazamiento>, "Mover") donde la acción se obtiene de un comando de voz y el parámetro <dirección desplazamiento> se obtiene del gesto realizado por el usuario.

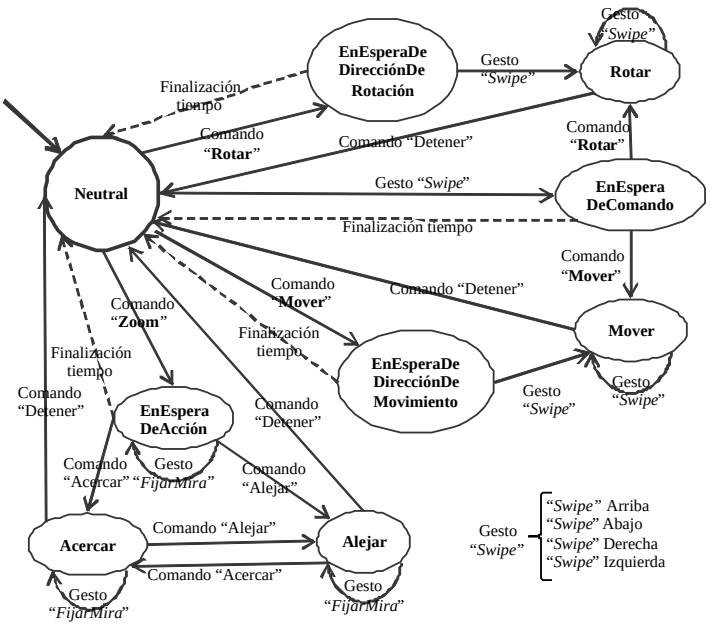

Fig. 7. Máquina de estados finitos describiendo la interacción multimodal con la aplicación Bing-Maps – Estado Inicial Neutral.

Rotación: requiere enunciar el comando de voz "Rotar" para llegar al estado "EnEsperaDeDirecciónDeRotación". Posteriormente se debe realizar el gesto de *swipe* en la dirección en que se quiera realizar la rotación (Figura 2). Esto generará que la visualización rote en la dirección deseada hasta que el usuario enuncie el comando "Detener". Es posible cambiar la dirección de la rotación haciendo el gesto *swipe* en la nueva dirección. Para esta acción se necesita la estructura ("Rotar", <dirección rotación>) o (<dirección rotación>, "Rotar") donde la acción se obtiene de un comando de voz y el parámetro <dirección rotación> se obtiene del gesto realizado por el usuario.

Detención: se debe enunciar el comando de voz "Detener" para detener la acción actual y llegar a un estado neutral donde la aplicación permite nuevamente las operaciones posibles de interacción.

En el estado neutral, las acciones Movimiento o Rotación pueden iniciarse también a partir del gesto *swipe* (Figura 2) pasando al estado "EnEsperaDeComando". Si en este estado se enuncian los comandos de voz "Mover" ó "Rotar" se llegará al estado "Mover" ó "Rotar" permitiendo desplazar o rotar la visualización en la dirección indicada por el gesto.

6. Evaluación y Resultados

La evaluación de la componente multimodal utilizada en el caso de estudio se basa en el método SUXES [17] para obtener una realimentación subjetiva del sistema multimodal. Se definen seis preguntas que los usuarios deben responder para tener una apreciación de su expectativa antes del uso del prototipo y de su experiencia de usuario después de realizar las tareas propuestas usando el prototipo de prueba. Las preguntas cuestionan aspectos específicos sobre las modalidades de voz y gestos que se usan en el prototipo multimodal con valores de apreciación de 1 (valor mínimo), 2, 3, 4 o 5 (valor máximo). Se completa la experiencia de usuario con cuatro preguntas apreciativas en la misma escala sobre el tiempo de respuesta del prototipo, su facilidad de uso, el nivel de agrado al usar la aplicación multimodal y si la interacción multimodal es adecuada para el caso de estudio.

6.1. Tareas de Evaluación

Las tareas de uso del prototipo multimodal de prueba que deben resolver los usuarios consisten en visitar las ciudades de Bogotá y París. Se considera que el usuario visitó una ciudad cuando esté dentro de la ciudad en el máximo nivel de zoom. La información de la

evaluación por cada usuario se recauda a través de un cuestionario. La primera parte del cuestionario tiene una parte de información general para caracterizar los usuarios participantes (edad, género y experiencia de uso de aplicaciones de mapas como Google Earth o Bing-Maps). La segunda parte del cuestionario corresponde a la expectativa (antes del uso) y la última parte de cuestionario corresponde a los resultados de su experiencia de usuario (después del uso). Por su parte, una persona participante del grupo de desarrollo del prototipo cumple el rol de evaluador y completa un cuestionario con medidas cuantitativas de desempeño de la componente multimodal para reconocer la intención de los usuarios en la realización de sus tareas: tiempo de realización de la visita a cada ciudad, número de aciertos y desaciertos en la detección de comandos de voz, número de aciertos y desaciertos en la detección de gestos.

6.2. Resultados

El estudio se realizó con nueve usuarios de edades entre 23 y 35 años con edad promedio de 28.6 años y niveles de estudios entre pregrado y postgrado, 3 mujeres y 6 hombres. Con respecto al nivel de expectativa, se encontró que los usuarios que diariamente usan aplicaciones de mapas tienen una expectativa más alta de la versión multimodal que los usuarios que usan aplicaciones de mapas con una menor frecuencia.

A continuación se presentan las preguntas comunes de Expectativa y Experiencia de usuario realizadas con respecto a las operaciones a realizar en la tarea de navegación de mapas y los resultados comparativos de los promedios para cada una de ellas (Figura 8):

- Pregunta común 1: ¿Considera Usted que el comando de voz para realizar el movimiento de traslación es el adecuado?
- Pregunta común 2: ¿Considera Usted que el comando de voz para realizar la rotación es el adecuado?
- Pregunta común 3: ¿Considera Usted que el comando de voz para realizar zoom es el adecuado?
- Pregunta común 4: ¿Considera Usted que los gestos para indicar la dirección del movimiento son los adecuados?
- Pregunta común 5: ¿Considera Usted que los gestos para indicar la dirección de la rotación son los adecuados?
- Pregunta común 6: ¿Considera Usted que el gesto para controlar la posición del zoom es el adecuado?

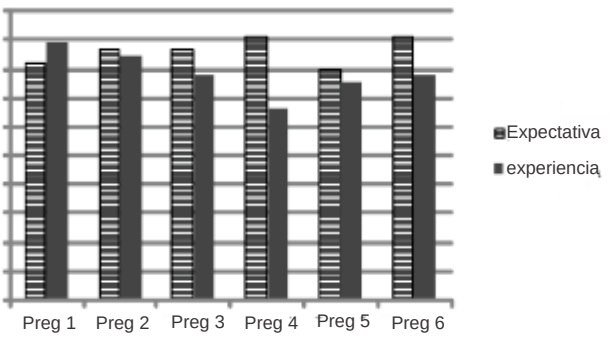

Fig. 8. Comparativo del valor promedio de las preguntas comunes de Expectativa vs Experiencia de usuario (Eje Y) detallado por pregunta (Eje X) para la población de muestra.

A partir de los resultados anteriores a las preguntas comunes 1 – 6, se obtuvo un promedio de la Expectativa por usuario de 25.9/30 y un promedio de la Experiencia por usuario de 23.6/30. Los usuarios proveen una calificación mayor en la expectativa que en la evaluación de experiencia para el prototipo multimodal de prueba.

Los usuarios completaron la tarea de llegar a París más rápido que a Bogotá usando la interfaz multimodal, lo que demuestra que al repetir la tarea se va incrementando la rapidez con que se realiza. El promedio para visitar Bogotá fue de 126 seg. mientras que para visitar París fue de 72 seg.

La efectividad del componente de reconocimiento de comandos de voz fue de 70%. La calificación promedio para los comandos de voz definiendo acciones de movimiento fue (Figura 9): 4.4/5.0 comando "mover"; 4.2/5.0 comando "rotar" y 3.9/5.0 comando "zoom". La efectividad del componente de reconocimiento de gestos fue de 74%. La calificación promedio para los gestos definiendo acciones de movimiento fue (Figura 10): 3.3/5.0 gesto *swipe* para mover; 3.8/5.0 gesto *swipe* para rotar y 3.9/5.0 gesto "FijarMira" (acción Zoom).

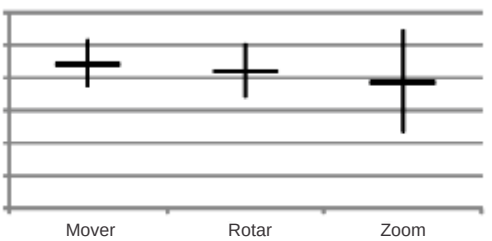

Fig. 9. Calificación promedio (línea horizontal) y desviación estándar (rango vertical) de los comandos de voz para la población de muestra.

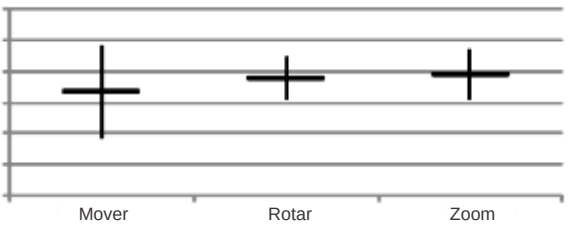

Fig. 10. Calificación promedio (línea horizontal) y desviación estándar (rango vertical) de los gestos para la población de muestra.

A continuación las preguntas apreciativas relacionadas con la experiencia de usuario en el prototipo de prueba y los promedios obtenidos a cada una de ellas (Figura 11):

- Pregunta 7: "¿El tiempo de respuesta del sistema multimodal de navegación de mapas es el adecuado?" – Calificación promedio: 3.8/5.0
- Pregunta 8: "¿Considera que fue fácil usar el sistema multimodal de navegación de mapas?" – Calificación promedio: 3.8/5.0
- Pregunta 9: "¿Su experiencia con el sistema multimodal de navegación de mapas fue agradable?" – Calificación promedio: 4.1/5.0
- Pregunta 10: "¿Considera que el mecanismo propuesto de multimodalidad (combinación gestos manuales-comandos de voz) es adecuado para la tarea de navegación en mapas?" – Calificación promedio: 4.0/5.0

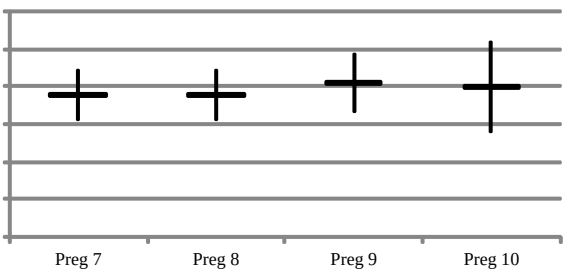

Fig. 11. Calificación promedio (línea horizontal) y desviación estándar (rango vertical) a las preguntas apreciativas para la población de muestra.

Los resultados más bajos corresponden a las preguntas 7 y 8 los cuales se dan probablemente porque el sistema multimodal está en desarrollo y aún presenta algunos problemas para el reconocimiento de gestos y de voz. Los resultados más altos corresponden a las preguntas 9 y 10 donde

se pregunta sobre la percepción de los usuarios al uso de la interfaz de interacción multimodal en la navegación de mapas.

7. Conclusiones y Trabajo Futuro

La interacción multimodal es de interés porque es natural para los usuarios lo cual favorece su experiencia al interactuar con una aplicación computacional. En particular, la combinación de gestos y voz ha mostrado que se complementan para definir acciones de navegación en aplicaciones visuales 2D/3D.

En este trabajo se propone una componente de software que permite el desarrollo de interfaces de interacción multimodales. La componente propuesta facilita la integración de múltiples modalidades y el modelado de la interacción de usuario usando el enfoque de máquinas de estados finitos. La componente se integró en un prototipo funcional que usa interacción multimodal basada en voz y gestos con las manos para la navegación en mapas 2D usando la aplicación Bing-Maps.

Después de realizar un estudio inicial con usuarios se concluye que el sistema con interacción multimodal genera bastante interés y expectativa en los usuarios y la experiencia de usuario se afecta por la calidad y desempeño en el reconocimiento de gestos y comandos de voz. Además, los resultados permitieron definir las direcciones del trabajo futuro:

- Mejorar la eficiencia de los módulos de reconocimiento aumentando la tasa de reconocimiento de gestos y comandos de voz y/o disminuyendo el tiempo de respuesta medido a partir de las acciones dadas por el usuario. Estas mejoras buscan aumentar la usabilidad y satisfacción del usuario al interactuar con una aplicación visual 2D/3D a través del uso de la interfaz multimodal.
- Automatizar la transformación del modelo de interacción multimodal (representado bajo una máquina de estados finitos) de una aplicación visual 2D/3D en la componente de software que define la interfaz de usuario multimodal y que se comunica con la aplicación.
- Proveer un mecanismo de aprendizaje de nuevos gestos y de nuevos comandos de voz buscando favorecer la extensibilidad y el desempeño de la componente de reconocimiento.
- Explorar el enfoque de correspondencia de patrones en el que se compara los movimientos del usuario con una base de datos de imágenes con los gestos a detectar[13].

- Probar la componente de interacción multimodal con otras aplicaciones de computación visual 2D/3D. Inicialmente se han identificado los casos de aplicaciones en visualización/análisis de imágenes para diagnóstico médico, video-juegos y/o en tele-operación de robots móviles en ambientes riesgosos.
- Mejorar el espacio de interacción teniendo en cuenta:
  - Modo Activación: Iniciar la interacción con el usuario con un gesto de inicio (*e.g.* gesto saludo/*wave*). Esto para asegurarse de que en el entorno puedan estar otras personas sin intervenir en la interacción.
  - Modo Desactivación: Interrumpir temporalmente la interacción con el usuario por atención de otras actividades o no estar en posición para ser detectado adecuadamente.
  - Identificar la dirección de la que provienen los comandos de voz para recibir únicamente comandos del usuario activo.

Referencias

[1] S. Oviatt, "Multimodal Interfaces," Handbook of Human-Computer Interaction, 2002, Capítulo 14, pp. 286-304.

[2] C. Duque, F. De la Rosa, J.T. Hernández. "Multimodal Interaction Architecture applied to Navigation in Maps," VIII Congreso Colombiano de Computación (8CCC), Armenia, Colombia, 2013.

[3] Microsoft Bing-Maps. (2013, Mar 4). [En línea] Disponible: http://www.bing.com/maps/.

[4] M.J. Egenhofer, "Query processing in spatial-query-by-sketch," *J. Visual Languages and Computing*, vol. 8, no. 4, pp. 403-434, 1997.

[5] D. McGee, P. Cohen, L. Wu, "Something from nothing: Augmenting a paper-based work practice via multimodal interaction," en *Proc. ACM Designing Augmented Reality Environments*, 2000, pp. 71-80.

[6] J. Florence, K. Hornsby, M.J. Egenhofer, "The GIS WallBoard: Interactions with spatial information on large-scale displays," en *7th Int. Symp. on Spatial Data Handling SDH*, 1996, pp. 449-463.

[7] M. Johnston, S. Bangalore, G. Vasireddy, "MATCH: Multimodal access to city help," en *Autom. Speech Recognit. and Understand. Works.*, 2001.

[8] L. B. Larsen, R. B. Jensen, K. L. Jensen, S. Larsen, "Development of an automatic pool trainer," en *Proc. ACM SIGCHI Int. Conf. of Advances in Computer Entertainment Technology ACE*, 2005, pp. 83-87.

[9] L.R. Ruiz, F. De la Rosa, J.T. Hernández, "Platform integrating interactive applications with gesture-based interaction," en *XXXVIII Conf. Latinoamericana en Informática CLEI*, 2012.

[10] Y. Ying, "A hierarchical approach to continuous gesture analysis for natural multi-modal interaction," en *Proc. 14th ACM Int. Conf. on Multimodal Interaction ICMI*, 2012, pp. 357-360.

[11] F. Daiber, S. Gehring, M. Löchtefeld, A. Krüger, "TouchPosing: multi-modal interaction with geospatial data," en *Proc. 11th Int. Conf. on Mobile and Ubiquitous Multimedia (MUM)*, 2012, Article 8.

[12] Microsoft. "Kinect for Windows SDK". (2013, Mar 4). [En línea] Disponible: http://msdn.microsoft.com/en-us/library/hh855347.aspx.

[13] D. Catuhe, *Programming with the Kinect for Windows software development kit*. Ed. Microsoft Press, 2012.

[14] Microsoft. "Microsoft Speech Platform". (2013, Mar 4). [En línea] Disponible: http://msdn.microsoft.com/en-us/library/hh361572.aspx.

[15] S. Oviatt, A. DeAngeli, K. Kuhn, "Integration and synchronization of input modes during multimodal human-computer interaction," en *Proc. ACM SIGCHI CHI Conference*, 1997, pp. 415-422.

[16] J. Chang, M.-L. Bourguet, "Usability framework for the design and evaluation of multimodal interaction," en *Proc. of the 22nd British HCI Group Annu. Conf. on People and Computers: Culture, Creativity, Interaction BCS-HCI*, 2008, pp. 123-126.

[17] M. Turunen, J. Hakulinen, A. Melto, T. Heimonen, T. Laivo, J. Hella, "SUXES-user experience evaluation method for spoken and multimodal interaction," en *Proc. of INTERSPEECH Conf.*, 2009, pp. 2567-2570.