Kinect. Innovación en IA

Competencia Transversal de Espíritu Emprendedor e Innovación

25/05/2013

Grado en Ingeniería Informática - UPC

Maria Claver
Borja González
Sergio Martínez

Tabla de contenido

1 Introducción

En el presente documento presentamos nuestro trabajo de investigación sobre un tema de innovación en Inteligencia Artificial. Como tema de estudio hemos escogido la tecnología de Kinect, producto de Microsoft para complementar la consola Xbox. Este producto es capaz de reconocer voz e imagen para entender órdenes y movimientos de los jugadores para que los juegos sean más interactivos y realistas.

Este trabajo consiste en el análisis de las técnicas de Inteligencia Artificial aplicadas en la tecnología de Kinect y la importancia que tienen en el desarrollo, el uso y la popularidad del producto. Veremos cuál ha sido la parte de la Inteligencia Artificial en este producto tecnológicamente innovador.

2 Kinect

Kinect es un dispositivo periférico para consolas Xbox y para PC's con Windows que reconoce las partes del cuerpo de los jugadores, detecta el movimiento y también incorpora reconocimiento facial y de voz. Permite la interacción con la consola o el ordenador sin necesidad de mandos ni contacto físico con ninguna interfície.

El dispositivo tiene el aspecto de la figura y se debe situar encima o debajo de la pantalla de juego en posición horizontal.

Figura 1: El dispositivo Kinect.

Su sensor principal es una cámara de profundidad, que, junto con un proyector de infrarrojos, permite captar la distancia a la que se encuentra cada objeto de la cámara. La cámara de la versión actual de Kinect tiene una resolución de 640x480 píxeles con 11 bits (2048 niveles de profundidad diferenciables). La nueva Kinect 2 será en cambio de alta resolución (1080p).

Además, cuenta con un sistema de microfonía y un software de reconocimiento de voz para captar la voz de los jugadores y atender a sus comandos. Los micrófonos están dispuestos de forma multi-vectorial de modo que se puede localizar la fuente del sonido y suprimir el ruido ambiente. La nueva consola Xbox (a la venta a finales de este año) incorporará la nueva versión de Kinect, que mantendrá el sistema de microfonía permanentemente encendido y tendrá la capacidad de "despertar mediante voz". De hecho, la consola sólo funcionará con el dispositivo Kinect conectado.

El periférico también cuenta con una cámara RGB estándar para captar las imágenes con luz y color, que se usa para el reconocimiento facial de los jugadores y la representación gráfica de éstos en dentro de los juegos.

La complejidad del sistema de Kinect radica en la mezcla de innovación en hardware y software, como muestra el uso de la cámara de profundidad y el sistema de reconocimiento de partes del cuerpo en 3D, o la matriz de micrófonos y la localización de las fuentes de sonido.

3 Técnicas de IA que utiliza

El dispositivo Kinect tiene varias funcionalidades, ya mencionadas, todas ellas complejas e interesantes, pero en este documento nos hemos centrado en la capacidad de detección de partes del cuerpo humano en tres dimensiones, por ser la más innovadora y la más destacada del dispositivo. En el reconocimiento de voz y el reconocimiento facial en 2D sabemos que también se aplican técnicas de inteligencia artificial, pero el uso de éstas no es nuevo y el producto que estudiamos no ha realizado un avance significativo en esta tecnología, así que no lo discutiremos aquí.

Para realizar el seguimiento del movimiento de un jugador, el dispositivo capta imágenes con la cámara de profundidad. Al funcionar con rayos infrarrojos, las imágenes grabadas con la cámara de profundidad no tienen en cuenta color ni luminosidad, evitando así los problemas que ocasionan en los sistemas convencionales las texturas y las sombras.

Figura 2: Imagen captada por una cámara de profundidad. En este caso, blanco es la profundidad menor y negro es la profundidad mayor detectada por la cámara.

De estas imágenes, como la que se muestra en la figura anterior, se obtiene la situación de las articulaciones de los jugadores. Esto se hace con algunos pasos intermedios. El primer paso intermedio, que Microsoft introduce como novedad, es el de obtener una "imagen" (o mapa) con las partes del cuerpo de los jugadores marcadas con colores como en la segunda imagen. Cada color se corresponde con una parte del cuerpo previamente definida.

Una vez que se tiene el paso intermedio de las partes del cuerpo, en forma de distribuciones probabilísticas, se calcula mediante relaciones aprendidas dónde está la articulación de cada parte del cuerpo detectada según dónde se haya calculado que está su centro de masas y se les añade la profundidad z que corresponde. Se obtienen así una serie de propuestas, acompañadas de su grado de fiabilidad, para la representación de las posiciones de las articulaciones del jugador. De cada propuesta se obtiene una visión frontal, una lateral y una superior para conformar la representación en tres dimensiones, como se puede ver en la figura siguiente.

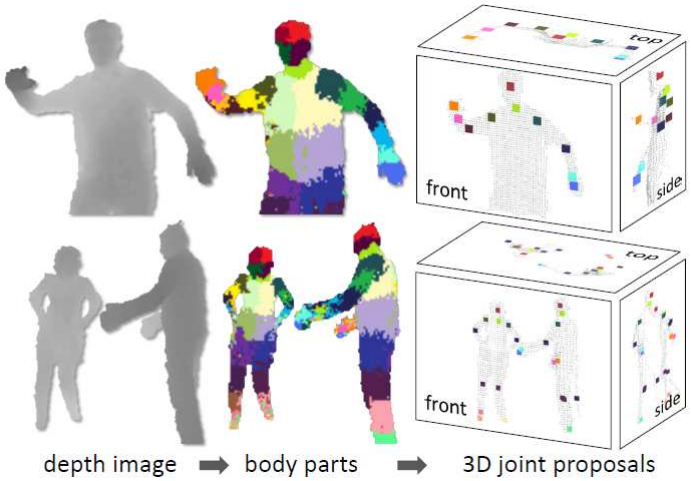

Figura 3: Paso de la imagen de la cámara de profundidad al mapa de probabilidades de partes del cuerpo y por último a una de las propuestas de las posiciones de las articulaciones en 3 dimensiones.

La parte interesante del proceso es la deducción de la pose del jugador a partir de una imagen. Este proceso se lleva a cabo con técnicas de aprendizaje automático (machine *learning*) con entrenamiento exhaustivo evitando el overfitting. En concreto, se utilizan bosques de decisión aleatorios (*random decision forests*) para la primera parte del proceso, que consiste en clasificar los píxeles de la imagen de profundidades según la parte del cuerpo a la que pertenecen, con una cierta distribución de probabilidades (*body part recognition*).

Veremos en el siguiente apartado qué son los bosques de decisión aleatorios y cómo se usan para hacer la clasificación de píxeles en partes del cuerpo.

4 Uso de las Técnicas de IA

Como ya hemos mencionado antes, Kinect usa los bosques de decisión aleatorios para clasificar los píxeles de las imágenes tomadas por la cámara de profundidad según la parte del cuerpo que pertenecen. Se obtiene entonces una imagen intermedia a partir de la original de profundidad con cada píxel pintado del color correspondiente a la parte del cuerpo a la que pertenece con más probabilidad, como se ve en la figura siguiente.

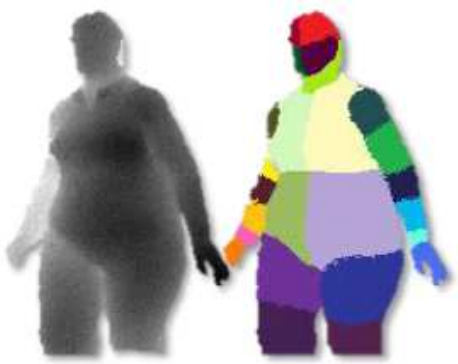

Figura 4: Paso de la imagen de profundidad, donde el negro es más cercano y el blanco es más lejano, al mapa de partes del cuerpo por píxel según probabilidad.

El proceso de clasificación es eficiente y relativamente sencillo, puesto que sólo se tiene que hacer entrar cada píxel por los árboles de decisión y en las hojas estará la distribución de probabilidad de la parte del cuerpo a la que pertenece. La dificultad estriba en el proceso de entrenamiento del sistema, que es cuando se construyen los árboles. Pero veamos cada cosa a su tiempo.

¿Qué es un bosque de decisión aleatorio?

Un bosque de decisión no es más que un conjunto de árboles de decisión usados en conjunto.

¿Qué es un árbol de decisión?

Un árbol de decisión es una estructura en forma de árbol que se hace recorrer a una entrada dada (que queremos clasificar) de la raíz hasta las hojas, pasando por los nodos de decisión interiores donde se evalúan las características que

definen a las distintas clases a las que puede pertenecer la entrada. En concreto, el sistema Kinect utiliza árboles de decisión binarios, es decir, cada nodo tiene dos hijos: sí o no.

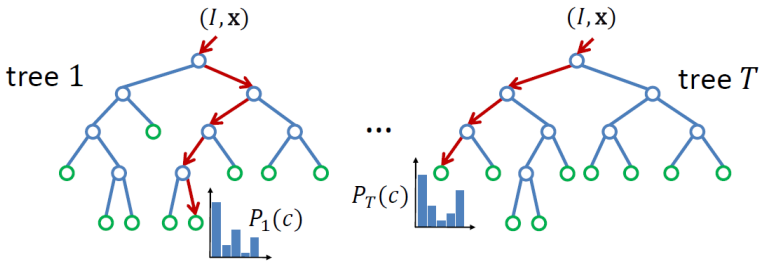

Figura 5: Bosque de decisión formado por T árboles de decisión. Los nodos de decisión son azules, las hojas con las distribuciones de probabilidad son verdes y el camino seguido por el píxel x de la imagen I en cada uno está marcado en rojo.

En nuestro caso, queremos clasificar píxeles de una imagen de profundidades, y las clases a las que puede pertenecer un píxel son las diferentes partes del cuerpo. El sistema Kinect distingue 31 partes distintas, distinguiendo izquierda y derecha, parte superior e inferior, etc. y también algunas articulaciones.

Se clasifican los píxeles individualmente, pero se quiere evaluar el resultado global de la imagen (si un píxel está mal clasificado, pero los de su alrededor bien, no es problema), así que la entrada del árbol será un píxel correspondiente a una imagen. En cada nodo interior del árbol, se evaluará una función que dirá si el píxel corresponde a una parte del cuerpo o no según el resultado. Estas funciones en general se llaman *weak learners*, y en el caso del algoritmo de Kinect evalúan las diferencias de profundidad según la imagen entre el píxel a clasificar y unos píxeles "cercanos" determinados.

Así, cada nodo mirará unos píxeles concretos (en posición relativa al píxel que miramos), por ejemplo, una función podría ser mirar si un píxel a unos 300px por encima tiene una profundidad un 50% menor que el que miramos. Si es que sí, podría tratarse de la parte superior del cuerpo, y si es que no, parece la parte inferior. La posición relativa al píxel en concreto es un parámetro inherente a la función, mientras que el 50% que nos hemos inventado como ejemplo es lo que se llama un umbral (*threshold)*, y veremos más adelante que se determina en el entrenamiento del árbol.

Las hojas en este caso son histogramas de probabilidades de pertenecer el píxel a cada una de las partes del cuerpo. En cuanto el píxel analizado llega a una hoja del árbol, se le asigna esa distribución de probabilidades. La parte del cuerpo en donde tenga el máximo será la que se le asigne al pintarlo y crear el mapa de probabilidades, pero el algoritmo no olvidará el resto de probables partes del cuerpo al hacer las propuestas de articulaciones en 3D.

¿Por qué el bosque de decisión es aleatorio?

Construir un árbol de decisión buscando el óptimo en todos sus parámetros (el subconjunto de funciones a evaluar, la distribución de las funciones en los nodos, los umbrales para cada función) es demasiado complejo y costoso cuando hablamos de dominios tan amplios como el que estamos analizando.

Por eso, se construyen varios árboles de decisión aleatorios, y se aumenta la fiabilidad de la clasificación haciendo pasar el elemento a clasificar por todos ellos, que es más rápido y eficiente que hacer un solo árbol con más profundidad. Además, se ha observado que la fiabilidad aumenta con el número de árboles en el bosque, mientras que disminuye a partir de cierta profundidad de los árboles, debido al overfitting que se produce. Para evitar el overfitting en un bosque de decisión, cada árbol se entrena con imágenes distintas.

Remarquemos que la aleatoriedad sólo aparece en el proceso de construcción del árbol, y que una vez está entrenado, el comportamiento de éste es determinista. Es decir, una vez construido el árbol, al hacer pasar un píxel de una imagen dada por el árbol, siempre acabará en la misma hoja. El proceso de construcción y entrenamiento los discutimos más adelante.

¿Qué es el overfitting?

El *overfitting* es el fenómeno que se da al dar demasiada información realista a un sistema de aprendizaje automático. Al tener demasiadas posibles respuestas para una pregunta aprendida, muchas veces solapadas, la fiabilidad de las respuestas a preguntas no conocidas disminuye.

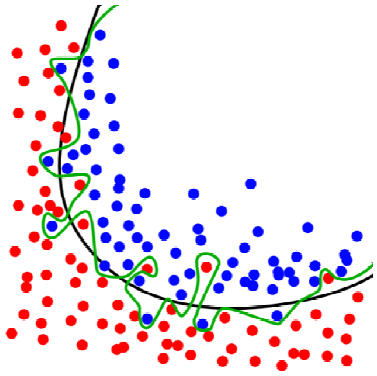

Figura 6: Overfitting. Al tener en cuenta todos los puntos, la máquina aprende el camino verde, que hará que tenga más errores en las predicciones.

¿Cómo y cuándo se construye el bosque?

El bosque de decisión aleatorio debe pasar una fase de entrenamiento, que consiste en la construcción de los árboles, para luego poder ser usado en lo que se llama la fase de "prueba", que es la fase en la que se puede usar para clasificar datos no conocidos.

Microsoft Research dicen tardar un día entero en entrenar un bosque de tres árboles de 20 niveles de profundidad con un millón de imágenes en un sistema distribuido de 1000 cores trabajando en paralelo, así que podemos hacernos una idea del enorme coste computacional de este proceso.

¿Cómo se entrena el sistema?

Para realizar el entrenamiento del bosque de decisión aleatorio, se necesitan imágenes de profundidad como entrada. Microsoft trabaja en un principio con fotografías sintetizadas con métodos de creación de gráficos por ordenador, imitando las imágenes que capta una cámara de profundidad, y en un momento más avanzado del proceso de entrenamiento utiliza también imágenes reales.

De cada imagen, se escogen 2000 píxeles de forma aleatoria. Cada uno de estos píxeles servirá de entrada a un árbol concreto, pero los píxeles de una imagen entrarán al mismo árbol, ya que queremos obtener resultados coherentes para

las imágenes enteras. Cada árbol se entrena con imágenes diferentes, para evitar el overfitting.

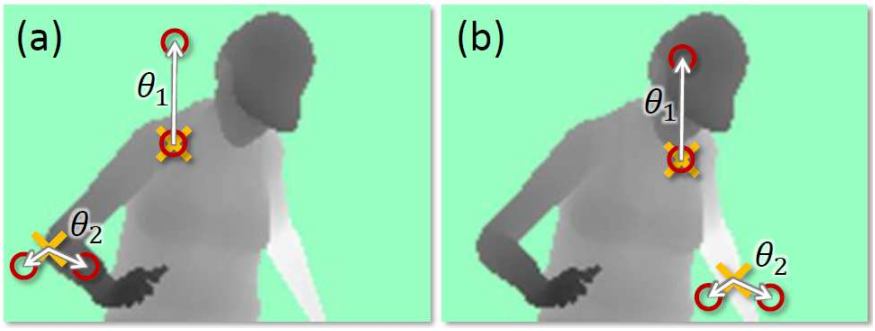

Figura 7: Dos características diferentes aplicadas a dos píxeles diferentes en a) y a otros dos en b), con distintos resultados.

Los nodos de los árboles se irán construyendo a medida que se le entren píxeles. Recordemos que en cada nodo tiene que haber una función que evalúe una característica del píxel y un umbral que indique si el píxel cumple o no la característica. Partimos pues de un conjunto (posiblemente infinito) de características evaluables, y un conjunto también grande de umbrales distintos para cada característica. Para la construcción de un nodo, se escoge un subconjunto aleatorio de estas características, y también un subconjunto aleatorio de sus posibles umbrales. Para sus experimentos, Microsoft Research usa conjuntos de 2000 características y 50 umbrales para cada una. Se evalúa entonces el píxel de entrada en cada una de esas combinaciones, y el que haya dado como resultado una mayor ganancia de información para el conjunto de la imagen (no para el píxel aislado), es la combinación que conformará el nodo.

Así se repite la operación para cada nodo y al final se llega a las hojas del árbol. Las hojas son la respuesta de la clasificación, son cada una un histograma de probabilidades de pertenencia a las diferentes partes del cuerpo posibles, como se ve en la Figura 5.

Cabe destacar que cuanto menor sea el tamaño del subconjunto de características escogidas para entrenar un árbol con respecto del número total de características, mayor será la aleatoriedad del árbol y menor la correlación entre los árboles de bosque, aumentando así su independencia y mejorando los resultados probabilísticos, además de minimizar el efecto del ruido en los datos.

5 Innovación e Impacto

La primera versión de Kinect se puso a la venta en noviembre de 2010 como periférico para la consola Xbox 360. Dos meses después, sus creadores reclamaron Récord Guiness por ser el dispositivo electrónico de consumo de venta más rapida, después de haber vendido más de 8 millones de unidades en 60 días.

En enero de 2011 se publicó el SDK para Windows 7 para que los desarrolladores pudieran hacer aplicaciones para Kinect en C++, C# o Visual Basic .NET. En febrero de ese mismo año se lanzó la versión de Kinect para Windows.

En 2011, Microsoft Research fue galardonada con el premio MacRobert de innovación, presentado por la Royal Academy of Engineering, por el trabajo de aprendizaje automático realizado con el proyecto Kinect de reconocimiento de movimiento humano. Asimismo, Kinect quedó en segundo lugar en la lista de los 10 productos más innovadores del 2011 de la revista Popular Mechanics. Microsoft sigue estando entre las 50 empresas más innovadoras en 2013 según el sitio web Fast Company.

Ahora está previsto el lanzamiento de la nueva versión del dispositivo, Kinect 2, que mejora la calidad del primero, aumenta la resolución de las cámaras y también la velocidad de procesado del movimiento. Además, incorpora la habilidad de despertar por orden de voz, ya que cuenta con un sistema de reconocimiento de voz y tendrá la microfonía siempre encendida. La primera versión de Kinect ya tenía la capacidad de reconocimiento de voz, aunque no en todos los países. La nueva Kinect 2 se espera saldrá al mercado junto con la consola Xbox One, que la llevará por defecto, hacia finales de 2013, a punto para la campaña navideña. La versión de Kinect 2 para Windows no se espera hasta 2014.

El reclamo de Kinect es el uso del lenguaje natural del jugador para interactuar con el juego. En ese campo, la competencia directa que tiene es con la consola Wii y su mando Wii Remote, que aunque requiere de mando para jugar,

interpreta bastante bien el movimiento del jugador que lo sostiene. También compite con Play Station Move y su control de movimiento de los ojos del jugador, aunque en menor medida.

Hasta la salida de Kinect al mercado, ninguno de los sistemas similares previos había conseguido funcionar a tiempo real, es decir, ninguno tenía una velocidad de procesado suficiente como para seguir los movimientos de un cuerpo entero, y menos aún generalizando para cualquier forma y tamaño de cuerpo.

La idea de la compañía al crear Kinect era conseguir llegar a una audiencia más amplia y variada que los consumidores estándar de la consola Xbox, pero lo que consiguieron superó con creces sus humildes expectativas. La tecnología de Kinect no solo sirve a los propósitos del entretenimiento, sino que su utilidad alcanza campos tan importantes como la medicina. Se conocen varios proyectos en los que se han adaptado dispositivos Kinect para ser usados con fines tan dispares como detectar signos de autismo en niños o servir de soporte en cirugía oncológica.

Desde luego, también se le encuentran otras utilidades menos ambiciosas y más divertidas, como es el uso del dispositivo para pasar presentaciones tipo PowerPoint o imitar la forma de interacción con ordenadores vista en la película Minority Report. Este tipo de aplicaciones permite ver un futuro de interacción con las máquinas mediante el lenguaje natural, de forma remota, sin necesidad de contacto físico. Quizá aplicaciones de soporte para discapacitados físicos es el siguiente paso, si no existe ya, o un paso más en la integración de los computadores en nuestro estilo de vida.

6 Bibliografía

1. Kinect. *Wikipedia.* [En línea] [Citado el: 24 de 05 de 2013.] http://en.wikipedia.org/wiki/Kinect.

2. Overfitting. *Wikipedia.* [En línea] [Citado el: 25 de 05 de 2013.] http://en.wikipedia.org/wiki/Overfitting.

3. **Microsoft Research.** Body Part Recognition. *Real-Time Human Pose Recognition in Parts from Single Depth Images.* [En línea] [Citado el: 24 de 05 de 13.] http://research.microsoft.com/pubs/145347/BodyPartRecognition.pdf.

4. **A. Criminisi, J. Shotton and E. Konukoglu.** Decision Forests. *Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning.* [En línea] [Citado el: 25 de 05 de 2013.] http://research.microsoft.com/pubs/155552/decisionForests_MSR_TR_2011_114.pdf.

5. Decision Tree Learning. *Wikipedia.* [En línea] [Citado el: 25 de 05 de 2013.] http://en.wikipedia.org/wiki/Decision_tree_learning.

6. Random forest. *Wikipedia.* [En línea] [Citado el: 25 de 05 de 2013.] http://en.wikipedia.org/wiki/Random_forest.

7. *Kinect's AI breakthrough explained.* [En línea] [Citado el: 25 de 05 de 2013.] http://www.i-programmer.info/news/105-artificial-intelligence/2176-kinects-ai-breakthrough-explained.html.

8. *Doctors use Xbox Kinect in cancer surgery.* [En línea] [Citado el: 25 de 05 de 2013.] http://chealth.canoe.ca/channel_health_news_details.asp?news_id=31715&news_channel_id=12&channel_id=12.

9. *Medical practice finds use for Kinect hack.* [En línea] [Citado el: 25 de 05 de 2013.] http://www.qj.net/qjnet/xbox-360/medical-practice-finds-use-for-kinect-hack.html.

10. *Minnesota University Team Adapts Kinect for Medical Use.* [En línea] [Citado el: 25 de 05 de 2013.] http://news.softpedia.com/news/Minnesota-University-Team-Adapts-Kinect-for-Medical-Use-189553.shtml.

11. *Control your PowerPoint and PDF presentations with Kinect.* [En línea] [Citado el: 25 de 05 de 2013.] http://www.prlog.org/11120062-control-your-powerpoint-and-pdf-presentations-with-kinect.html.

12. *Kinect finally fulfills its Minority Report destiny.* [En línea] [Citado el: 25 de 05 de 2013.] http://www.engadget.com/2010/12/09/kinect-finally-fulfills-its-minority-report-destiny-video/.

13. FastCompany. *Most innovative companies 2013.* [En línea] [Citado el: 26 de 05 de 2013.] http://www.fastcompany.com/section/most-innovative-companies-2013.

14. Popular Mechanics. *The 10 most innovative tech products of 2011.* [En línea] [Citado el: 26 de 05 de 2013.] http://www.popularmechanics.com/technology/gadgets/reviews/the-10-most-innovative-tech-products-of-2011#slide-1.