Reconocimiento de gestos aplicado al control de dispositivos

Una experiencia en control de TV

Carla Luna Gennari, César Estrebou, and Laura Lanzarini

Instituto de Investigación en Informática LIDI. Facultad de informática.
Universidad Nacional de La Plata
carla.lunagennari@gmail.com,{cesarest,laural}@lidi.info.unlp.edu.ar

Resumen El avance tecnológico ha permitido construir controladores cada vez más elaborados a la hora de desarrollar una interfaz hombre-máquina amigable. En esta dirección, el reconocimiento de gestos hechos con las manos ha recibido gran atención ya que su implementación sólo requiere de un hardware mínimo.
Este trabajo presenta el prototipo de un sistema reconocedor de gestos capaz de detectar las manos y realizar su seguimiento. Según el gesto del cual se trate se generan comandos que al transmitirlos de manera infrarroja permiten controlar un dispositivo electrónico. Durante el desarrollo se combinaron soluciones de distintas áreas entre las que se destacan: redes neuronales, procesamiento de imágenes y electrónica. Los resultados obtenidos han sido satisfactorios. Cabe aclarar que, si bien el desempeño de la solución obtenida ha sido medido al controlar las funciones básicas de un televisor, la misma puede ser aplicada en numerosas situaciones.

Keywords: Segmentación de Manos, Reconocimiento de Gestos, Interfaz Hombre-Máquina, Control de Dispositivos

1. Introducción

Una interfaz hombre-máquina juega un papel importante a la hora de transmitir una intención de un usuario a un dispositivo. A medida que la tecnología avanza se incorporan al mercado soluciones nuevas con distintos tipos de hardware y una forma de comunicación propia. Elementos convencionales como el mouse y teclado, hacen que la operatoria en ocasiones sea compleja. En cuanto a los comandos por voz, si bien han sido ampliamente utilizados, presentan serias limitaciones ante la diversidad de idiomas o cuando deben operar en ambientes ruidosos.

Por lo antes dicho, hoy en día existe mucho interés en el reconocimiento de gestos hechos con las manos. En la actualidad hay diferentes trabajos realizados con aplicación en áreas muy diferentes como por ejemplo realidad aumentada, control de dispositivos o reconocimiento del lenguaje de señas Argentino entre otras [6,7,11]. También existen trabajos relacionados donde se utilizan los gestos

hechos con las manos para controlar las funciones básicas de un equipo de audio ubicado dentro de un vehículo [9].

Para llevar a cabo el reconocimiento de gestos hechos con las manos deben resolverse dos partes perfectamente diferenciadas. La primera de ellas se refiere específicamente a la segmentación de las manos y la segunda se ocupa de la caracterización y reconocimiento del gesto. Ambas presentan diferentes niveles de complejidad según el tipo de cámara que se utilice así como la cantidad y el tipo de gestos a reconocer.

En lo que se refiere a la cámara de video, existen dispositivos que brindan información en 3D permitiendo captar no sólo el desplazamiento de los objetos sino la profundidad a la cual se encuentran con respecto el punto de observación [2]. Por ejemplo en [5] se utiliza un mapa de profundidad para ayudar a la segmentación. La elección de la cámara es una relación de compromiso entre el costo del equipamiento necesario para resolver el problema y la complejidad del algoritmo de segmentación a desarrollar. En este trabajo se ha decidido utilizar una cámara de video convencional e identificar la zona en la cual se encuentran las manos a través de un reconocedor de colores de piel basado en una red neuronal similar a la indicada en [12, 15]. Para llevar a cabo su entrenamiento se construyó una base de datos de colores de piel ya que al momento de diseñar este prototipo no se disponía de este tipo de información.

Para la segunda parte del problema es necesario tener en cuenta la diversidad de formas que puede tomar una mano humana a la hora de realizar un gesto. El reconocimiento está ligado a la elección especifica de los gestos que se quieren procesar.

Este trabajo presenta el prototipo de un sistema reconocedor de gestos hechos con una mano. Según el gesto del cual se trate, genera comandos que al transmitirlos de manera infrarroja permiten controlar un dispositivo electrónico. Se trata de una solución de bajo costo para la que se describe tanto el diseño de hardware como el software de control. En este caso particular, fue utilizado para reemplazar el control remoto de un TV, posibilitando cambiar de canal y subir o bajar el volumen con sólo mover una mano frente al televisor. Su diseño modular permite incorporar el reconocimiento de nuevos gestos así como los comandos necesarios para operar con otros aparatos.

Utilizar gestos hechos con las manos para interactuar con dispositivos puede resultar sumamente útil para personas con movilidad reducida.

2. Hardware para Captura y Control de dispositivos

Una parte fundamental en la construcción de este prototipo fue la elección de los componentes de hardware para obtener un dispositivo embebido que respondiera tanto a los requerimientos de software como a los de futuras ampliaciones de hardware. Los requerimientos propuestos que debía cumplir son los siguientes: bajo costo económico, capacidad de cómputo para procesamiento de imágenes, capacidad de comunicación con otros dispositivos a bajo nivel, tamaño reducido,

(a) (b) (c) (d)

Figura 1: Módulos que componen el hardware. (a) Raspberry PI 3. (b) Cámara web. (c) Sensor de luminosidad. (d) Emisor y receptor infrarrojos.

bajo consumo de energía, capacidad de funcionar con baterías y capacidad de actualización de hardware y software.

Teniendo en cuenta los requerimientos antes mencionados se decidió utilizar una Raspberry Pi 3 (Fig. 1a) como corazón del dispositivo. Cuenta con 1Gb de memoria RAM y un procesaror ARM Cortex-A53 de 1.2GHz con 4 núcleos lo que ofrece un nivel de cómputo aceptable para procesamiento de imágenes moderado. También tiene un tamaño reducido, capacidad de comunicación con otros dispositivos a través de su GPIO y de sus 4 puertos USB. Cuenta con comunicación inalámbrica vía WiFi y Bluetooth. Requiere una alimentación de 5V y 1.5A para nuestro tipo de aplicación, lo que hace posible el uso de baterías, como por ejemplo un powerbank. Todo esto a un costo razonable.

Para capturar los gestos de las manos se utilizó una cámara web convencional (Fig. 1b) con una tasa de adquisición de unos 20 cuadros por segundo a una resolución de 1280 x 1024 píxeles. Para controlar las funciones del TV se utilizó un módulo emisor de luz infrarroja (Fig. 1d) y un módulo receptor infrarrojo que permite incorporar al sistema los códigos de aquellos controles remotos que no tengan códigos conocidos. También se utilizó un módulo para sensar la luz ambiente (Fig. 1c) con el objetivo de realizar ajustes en la luminosidad de la imagen capturada con la cámara web.

3. Software para Reconocimiento de Gestos

Teniendo en cuenta que esta es la primera versión del dispositivo, se decidió limitar su funcionamiento a una aplicación relativamente simple que no demandara una gran cantidad de cómputo. De esta manera se construyó un prototipo funcional tanto de hardware como de software que permite reemplazar el control remoto de un televisor convencional por un controlador que reciba las instrucciones a través de gestos realizados con las manos. En este sentido se limita tanto el modelo de representación como la cantidad de gestos de la mano, aunque en futuras versiones se tiene planificado ir incorporando un modelo más complejo como el planteado aquí [10].

Respecto del software utilizado para el desarrollo de este prototipo, se utilizó la versión Jessie de Raspbian, que es el sistema operativo con soporte oficial

de Raspberry. El sistema fue implementado en Python por su facilidad para prototipado rápido y por la gran disponibilidad de bibliotecas que ofrecen tanto algoritmos tradicionales como de vanguardia. Para las partes que requirieron procesamiento de imágenes se utilizó la biblioteca OpenCV que tiene una sólida madurez y sus algoritmos están altamente optimizados, incluso para aprovechar las características de la GPU. Tanto para el entrenamiento y funcionamiento de la red neuronal RCE como para el algoritmo de seguimiento de la mano se utilizó una implementación propia optimizada para NumPy.

En las secciones que siguen a continuación se describen en detalle los pasos de cada una de las etapas en las que se divide todo el proceso de reconocimiento del gesto para controlar un TV. En el esquema de la figura 2 se puede observar todo el proceso que realiza el software del dispositivo.

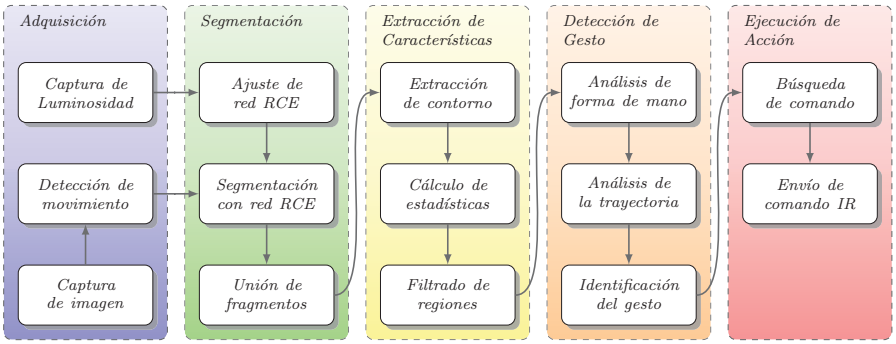

Figura 2: Proceso de reconocimiento del gesto y ejecución de comandos.

3.1. Adquisición

La etapa de adquisición se divide en dos partes. La primera consiste en la captura, mediante una cámara web convencional, de una imagen RGB con una resolución de 1280 x 1024 píxeles a una tasa de adquisición que puede variar de 10 a 20 cuadros por segundo. Una vez obtenida la imagen correspondiente al cuadro actual, se la compara con la del cuadro adquirido anteriormente para determinar si se produjo movimiento y así evitar el procesamiento innecesario.

La segunda parte consiste en la adquisición de la cantidad de luz ambiente a través de un transductor que mide esta variable en lux ($lumen/m^2$). Este valor obtenido permite ajustar la luminosidad de la imagen en la etapa de segmentación y de esta manera corregir las variaciones en la iluminación que afectan a los colores.

3.2. Segmentación

Para realizar la detección del color de piel en la imagen se utilizó una red neuronal RCE [8] (Restricted Coulomb Energy). Esta arquitectura de red presenta tres capas de neuronas, donde la primer capa y la segunda están completamente conectadas, y la segunda capa con la tercera solo estén conectadas parcialmente. La capa de entrada es decir la primera, es representada por los valores del espacio de color utilizado, en este caso RGB. La capa intermedia, llamada capa de prototipo, contiene información de color de los píxeles, los cuales corresponden a valores RGB que representan colores de piel que fueron incorporados durante el entrenamiento de la red o etapa de aprendizaje. La última capa o capa de salida contiene las clases a las que puede corresponder el color del pixel entrante. En este caso sólo hay una única clase que responde para determinar si un color corresponde o no a la piel humana.

Para realizar la segmentación de la piel humana se ingresa cada pixel de la imagen a la red RCE para determinar si éste se corresponde o no al color de la piel. Como resultado de esta operación se obtiene una máscara preliminar a la que luego se le aplica una operación morfológica de cierre para unir áreas que pudieran haber quedado desconectadas.

Es importante destacar que antes de realizar la segmentación, con el objetivo de subsanar el problema de inestabilidad que provocan las variaciones de iluminación, se aplica un ajuste a la red RCE según la luz ambiente captada por el sensor de luminosidad. Esta corrección se aplica a los valores RGB de las neuronas de la capa intermedia para neutralizar los cambios de intensidad de luz que están presentes en la imagen.

3.3. Extracción de características

En esta etapa, luego de obtenidas las áreas de la imagen donde se localiza la piel, se determina cuales de todas esas porciones pueden corresponderse con la mano de una persona.

El proceso inicia con la máscara que representa a todos los píxeles reconocidos como piel en la etapa anterior. A ésta máscara se le aplica un algoritmo de extracción de contornos para obtener un listado de estos. Cada contorno agrupa píxeles interconectados en la máscara que forman una región donde potencialmente podría haber una mano.

Una vez obtenido el listado, por cada contorno se calcula el área, el perímetro y los ejes principales que son propiedades que dan una pauta de las características geométricas generales de la región.

Finalmente, para filtrar el listado, se analiza cada contorno o región comparando sus propiedades con las propiedades de una región que contiene una mano, descartando aquellos que difieren mucho de lo esperado. Como resultado de esta comparación se obtiene un listado de contornos candidatos que representan regiones potenciales donde puede encontrarse una mano.

3.4. Reconocimiento del gesto

Para determinar si las regiones candidatas obtenidas en la etapa anterior se corresponden o no con una mano se aplica la técnica de *Template Matching* . En esta técnica se utiliza un listado de plantillas pre-definidas que representan las formas de las manos aceptadas o reconocidas por la aplicación. Cada región del listado de candidatas se compara con cada plantilla pre-definida para determinar el grado de correlación que existe entre ambas. La forma de la mano reconocida es la que mayor valor de correlación tenga, siempre que se supere un umbral establecido como parámetro de configuración.

Una vez que se obtiene una región candidata que coincide con alguna de las formas pre-definidas de mano, se inicia el proceso de seguimiento tomando como referencia la posición de la región como posición de inicio. Luego se analiza la secuencia posterior de imágenes localizando la mano para determinar la evolución de su posición. Cuando la distancia entre la posición inicial y la posición actual de la mano supera un valor determinado por un parámetro de configuración se establece la dirección del movimiento y se procede a determinar si el gesto coincide con los definidos en la aplicación.

3.5. Ejecución de Acción

Identificado el gesto de la mano, se determina el comando infrarrojo que éste tiene asociado para enviarlo al TV. Para ello, se utiliza un pequeño módulo de hardware que genera una señal infrarroja que es recibida por el TV, quien la decodifica e interpreta para ejecutar la función asociada a dicha señal.

Para codificar los comandos infrarrojos se utiliza la biblioteca LIRC (Linux Infrared Remote Control). Esta biblioteca permite decodificar y reproducir una secuencia de pulsos infrarrojos de la misma manera que lo hace un control remoto convencional. Cuenta con una gran cantidad de códigos predefinidos de controles remotos e incorpora un servicio con la capacidad de copiar, desde un control remoto, aquellos códigos que no están predefinidos.

(a) (b) (c)

Figura 3: Reconocimiento del gesto y acción realizada. (a) Subir el volumen. (b) Bajar el volumen. (c) Siguiente canal.

4. Pruebas/Resultados

4.1. Sistemas de Color y Red Neuronal RCE

Para la segmentación de manos se utilizó una red neuronal RCE [8] que determina cuando un pixel de la imagen se corresponde con el color de la piel. En la revisión de la literatura sobre la segmentación de piel basada en el color del pixel [1, 4, 13, 14] se encuentra que distintos algoritmos aplicados a imágenes en diferentes sistemas de representación del color obtienen resultados aceptables. Por este motivo, se decidió realizar una serie de pruebas en los sistemas de representación RGB, HSL, HSV, YCbCr, Cie-LAB para determinar cual es el más conveniente. Para realizar las pruebas en los distintos sistemas se utilizó la base de datos MOHI [3] que contiene muestras de manos de 250 personas y una base de datos construida ad hoc con imágenes que no contienen piel.

En la figura 4 se muestran los resultados obtenidos en las distintas pruebas realizadas. Para cada sistema de color se muestran dos barras que expresan el promedio de píxeles clasificados como piel cuando la imagen contiene piel y cuando no contiene piel. En general se puede observar que no hay grandes diferencias entre los distintos sistemas. También se puede observar que una mejora en la detección de los pixeles de la BDD con piel incrementa la cantidad de píxeles detectados en la BDD sin piel y viceversa. En consecuencia, se decidió optar por utilizar el sistema RGB para la segmentación con la red RCE para evitar el costo del cómputo de la transformación a otro sistema.

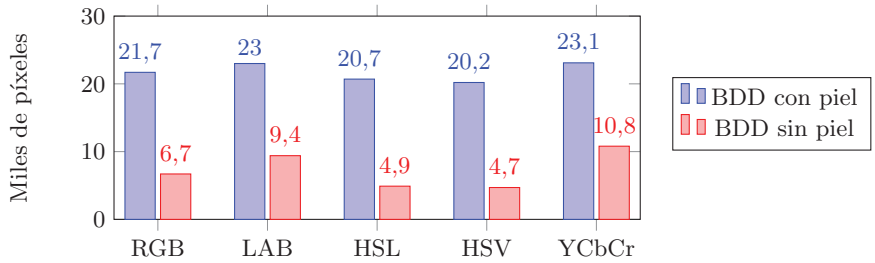

Figura 4: Red neuronal RCE en diferentes sistemas de representación del color.

4.2. Reconocimiento de Gestos

Con el prototipo terminado se realizaron pruebas para medir su precisión en un ambiente controlado. Para esto se utilizaron 11 sujetos en dos condiciones diferentes de iluminación. Una condición de iluminación es baja (30 lux) y la otra condición de iluminación es media (al menos unos 600 lux). De cada sujeto se tomaron 3 formas de la mano (figura 6): mano abierta con dedos separados, mano abierta con dedos juntos y mano cerrada con indice y pulgar separados.

Con cada forma se realizaron movimientos en dos direcciones opuestas (arriba y abajo). En la tabla de la figura 5 se pueden observar los resultados obtenidos.

Gesto	Distancia 180 cm		Distancia 300 cm	
	+600 lux	30 lux	+600 lux	30 lux
Mano Abierta - Dedos separados - Arriba	100 %	100 %	100 %	100 %
Mano Abierta - Dedos separados - Abajo	100 %	100 %	100 %	100 %
Mano Abierta - Dedos juntos - Arriba	100 %	45 %	100 %	33 %
Mano Abierta - Dedos juntos - Abajo	100 %	45 %	100 %	33 %
Mano con índice y pulgar - Arriba	40 %	0 %	30 %	0 %
Mano con índice y pulgar - Abajo	40 %	0 %	30 %	0 %

Figura 5: Resultados en porcentaje de la detección de los gestos.

(a) (b) (c)

Figura 6: Formas de manos procesadas. (a) Mano Abierta con dedos separados. (b) Mano abierta con dedos juntos. (c) Mano con índice y pulgar.

Para las pruebas con las formas de las manos se puede observar que para los casos de mano abierta se pueden detectar bien en condiciones de iluminación aceptable. Para los casos de mano abierta con dedos juntos se puede observar que falla cuando la iluminación es baja. Para los casos de la mano con índice y pulgar se observa que la detección falla de manera importante en las 2 condiciones de iluminación. La principal causa de este problema es la sombra que genera la flexión de los dedos. Esta sombra hace que la segmentación falle al no poder reconocerla como piel y se produzcan separaciones importantes que hacen que la región de la mano se extraiga parcialmente. Luego no se encuentran coincidencias o bien porque la región extraída se descarta porque no cumple con las propiedades geométricas esperadas o porque la técnica de *Template Matching* falla porque no es robusta para encontrar coincidencias parciales.

5. Conclusiones y Trabajos Futuros

En este trabajo se ha presentado la primera versión de un prototipo de un sistema reconocedor de gestos realizados con las manos para controlar las funciones de un TV.

Uno de los resultados positivos de este trabajo es que se ha obtenido una segmentación de piel aceptable utilizando una red neuronal RCE que funciona independientemente del sistema de color en que utiliza la imagen.

Un aspecto aún no resuelto es la sensibilidad del reconocedor propuesto a la variación de la luz ambiente. En este sentido, se encontró una solución parcial a esta dificultad utilizando lecturas de un sensor de luz para adaptar la red neuronal RCE a las condiciones de iluminación del ambiente. Cuando hay ausencia de luz ambiente o esta es notablemente baja, la corrección por intensidad de la luz no resulta suficiente para una cámara web convencional. En este caso es conveniente realizar una adaptación de la cámara para que funcione con iluminación infrarroja.

Debido a que este es un trabajo inicial, hay algunas líneas de trabajo que se están desarrollando y otras líneas que quedan por explorar y desarrollar. Actualmente se está trabajando para mejorar la segmentación de las áreas de la piel. Muchas veces debido a la combinación de iluminación con la posición de las manos, se producen sombras en las uniones de las falanges que provocan que la segmentación deje los dedos separados de las manos. Una solución que se está explorando es la incorporación al entrenamiento de la red neuronal RCE ejemplos de piel con sombra, lo que permitiría mejorar la segmentación en la unión de las falanges. Los resultados obtenidos en algunas pruebas preliminares que se realizaron al momento de la escritura de este documento son prometedores en este aspecto.

En lo que se refiere al hardware, aún resta explorar las maneras de controlar las funciones distintos dispositivos electrónicos mediante red cableada, red inalámbrica, bluetooth e infrarrojo para aprovechar las capacidades del dispositivo y expandirlas. Finalmente quedan por analizar algunas alternativas de reemplazo de la Raspberry Pi 3 para reducir el costo. Dispositivos como Orange Pi, en sus versiones Zero, One y Lite podrían reducir el costo entre un tercio y la mitad de una Raspberry. Estos dispositivos son compatibles y ofrecen características de hardware similares con algunas limitaciones que van desde una menor cantidad de memoria RAM, menos puertos USB hasta la ausencia de WiFi y bluetooth según el modelo.

Referencias

1. Chaves-González, J.M., Vega-Rodríguez, M.A., Gómez-Pulido, J.A., Sánchez-Pérez, J.M.: Detecting skin in face recognition systems: A colour spaces study. Digit. Signal Process. 20(3), 806–823 (May 2010)
2. Dinh, D.L., Kim, J.T., Kim, T.S.: Hand gesture recognition and interface via a depth imaging sensor for smart home appliances. Energy Procedia 62(Complete), 576–582 (2014)

3. Hassanat, A., Al-Awadi, M., Btoush, E., Al-Btoush, A., Alhasanat, E., Altarawneh, G.: New mobile phone and webcam hand images databases for personal authentication and identification. Procedia Manufacturing 3, 4060 – 4067 (2015), 6th International Conference on Applied Human Factors and Ergonomics and the Affiliated Conferences, AHFE 2015
4. Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A survey of skin-color modeling and detection methods. Pattern Recogn. 40(3), 1106–1122 (Mar 2007)
5. Kang, B., Tan, K., Tai, H., Tretter, D., Nguyen, T.Q.: Hand segmentation for hand-object interaction from depth map. CoRR abs/1603.02345 (2016)
6. Piumsomboon, T., Clark, A., Billinghurst, M., Cockburn, A.: User-defined gestures for augmented reality. In: CHI '13 Extended Abstracts on Human Factors in Computing Systems. pp. 955–960. CHI EA '13, ACM, New York, NY, USA (2013)
7. Reifinger, S., Wallhoff, F., Ablassmeier, M., Poitschke, T., Rigoll, G.: Static and dynamic hand-gesture recognition for augmented reality applications. In: HCI Intelligent Multimodal Interaction Environments: 12th International Conference, Part III. pp. 728–737. Springer, Berlin, Heidelberg (2007)
8. Reilly, D.L., Cooper, L.N., Elbaum, C.: A neural model for category learning. Biological Cybernetics 45(1), 35–41 (Aug 1982)
9. Riener, A.: Gestural interaction in vehicular applications. In: Computer. vol. 45, pp. 42–47. IEEE Computer Society, Los Alamitos, CA, USA (2012)
10. Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L.: Clasificaci'on de configuraciones de manos del lenguaje de se nas argentino con probsom. In: XXI Congreso Argentino de Ciencias de la Computacii'on (2015)
11. Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L., Rosete, A.: Sign Languague Recognition Without Frame-Sequencing Constraints: A Proof of Concept on the Argentinian Sign Language, pp. 338–349. Springer, Cham (2016)
12. Sui, C., Kwok, N.M., Ren, T.: A restricted coulomb energy (rce) neural network system for hand image segmentation. In: Computer and Robot Vision (CRV), 2011 Canadian Conference on. pp. 270–277. IEEE (2011)
13. Vezhnevets, V., Sazonov, V., Andreeva, A.: A survey on pixel-based skin color detection techniques. In: GraphiCon. pp. 85–92 (2003)
14. Xu, Z., Zhu, M.: Color-based skin detection: survey and evaluation. In: 2006 12th International Multi-Media Modelling Conference. pp. 10 pp.– (2006)
15. Yin, X., Guo, D., Xie, M.: Hand image segmentation using color and rce neural network. Robotics and Autonomous Systems 34(4), 235–250 (2001)