ISSN 1870-4069

Acceso automático por video detección de gestos manuales utilizando técnicas eficientes de segmentación dinámica de color de piel

Julio Zamora Esquivel, José Rodrigo Camacho Pérez, Cruz Vargas Jesus Adan, Paulo López Meyer, Héctor Cordourier Maruri

Intel Labs, GDC, Guadalajara, México
julio.c.zamora.esquivel@intel.com

Resumen. En este trabajo, proponemos un algoritmo ultra-ligero de video detección para la activación (o "wake up" en inglés) automática de dispositivos de cómputo. La propuesta le permite al usuario interactuar con el dispositivo utilizando gestos manuales que son detectados por video desde la perspectiva de primera persona.

Palabras clave: visión por computadora, reconocimiento de patrones, realidad aumentada.

Hand-Gesture Wake-Up Using Efficient Dynamic Skin Tone Segmentation

Abstract. In this work we propose an ultra lightweight visual-based method to wake-up a visual system, allowing the user to interact with the device using hand-gestures (i.e. combinations of hand poses) through first person view by means of a wearable camera.

Keywords: computer vision, pattern recognition, virtual reality, wake-up.

1. Introducción

Actualmente, los sistemas de interacción humano-computadora basados en la video detección de gestos manuales operan de manera continua. Es decir, el sistema captura video y simultáneamente opera algoritmos de detección de gestos de mano sin interrupción. Esto implica un alto costo energético, en particular al considerar la posibilidad de periodos de tiempo relativamente largos sin actividad de interacción por parte del usuario. Por ejemplo, en el caso de soluciones basadas en redes neuronales artificiales (RNA), esta necesidad de operación continua implica que el sistema analice todos los cuadros de video capturados, siendo evaluados por al RNA con diferentes resoluciones de la captura. Estas evaluaciones y su consecuente alto costo energético impactan en la experiencia de uso en la práctica, como por ejemplo, la reducción de vida de batería en el

caso de los dispositivos móviles. Alternativas a la operación continua incluyen el uso de botones de interacción, reconocimiento de frases clave habladas o algunas otras opciones que requieren de hardware adicional. Sin embargo, estas opciones tienen un impacto en la experiencia del usuario ya que demeritan la sensación de inmersión en los ambientes de realidad virtual. Algunas alternativas para solucionar este problema incluyen:

1. **Activación no visual**. Para evitar la operación continua del sistema de video reconocimiento, comúnmente se utilizan métodos de activación que requieren interacciones táctiles (pulsar botones en teclados o realizar gestos en pantallas o superficies táctiles) o vocalización de palabras o frases clave. Entre los ejemplos comerciales de solución de este tipo se tienen las teclas de activación en teléfonos móviles o las frases .ªlexaz .ºK Google"de los servicios de Amazon y Google respectivamente.
2. **Video detección de gestos de activación con Histogramas de Gradientes**. En este caso, se construye un descriptor por cada patrón basado en un Histograma de Gradientes de cada pixel en la imagen capturada y una RNA clasifica el descriptor. El gradiente es calculado por cada pixel en una imagen de tamaño $n \times m$ utilizando la convolución de la imagen con dos núcleos (o "kernels.[en] inglés).
3. **Video detección de gestos de activación con Aprendizaje Masivo (o "deep learning.[en] inglés**. En este enfoque, se realiza la segmentación del gesto manual con un pre-procesamiento de la imagen binarizada para reducir el tamaño del modelo de RNA. Esto es seguido por una sub-segmentación de cada bloque de color para generar sub-clases. Finalmente, el modelo de RNA es aplicado a cada pixel para detectar los gestos.

El método propuesto en este trabajo está basado en el tercer enfoque mencionado y es particularmente útil en el caso prendas electrónicas (o "wearables") de realidad virtual como el "HoloLens"para activar o re-activar periodos de bajo consumo de energía con las siguientes ventajas:

- Reducción de consumo de energía. Este método permite el uso de una sóla evaluación de la RNA por cada image en contraste con las $C \times O$ evaluaciones necesarias en un enfoque de Aprendizaje Masivo (o "deep learning"). Por ejemplo, si se tienen $C = 7$ colores y $O = 4$ objetos candidatos, el número total de evaluaciones necesarias por imagen capturada sería de 28, comparada con solo 1 en la presente propuesta.
- Reducción de falsos positivos. El método propuesto mejora la eficiencia de detección ya que compensa las tonalidades de la piel y las condiciones de iluminación en cada imagen capturada sin costo computacional agregado ya que se define una segmentación dinámica de color de piel.
- Mejora en la experiencia de uso. El método propuesto es completamente basado en video detección de gestos y por ende el usuario no necesita interrumpir la experiencia virtual ni pulsar botones o superficies táctiles.

2. Antecedentes

Existe una variedad de algoritmos de detección de gestos [2,3] los cuales ofrecen resultados confiables pero con un costo financiero, computacional y en facilidad de integración a dispositivos móviles o prendas electrónicas ("wearables"). Como se ha mencionado, algunas de estas soluciones se basan en el Reconocimiento de Claves Habladas (RCH), es decir el reconocimiento de palabras o frases específicas habladas ("Keyword Spotting Systems.$^{\text{en}}$ inglés), los cuales requieren de respuesta en tiempo real lo mismo que de exactitud para ofrecer experiencias de uso aceptables. Típicamente, estos sistemas están compuestos de un algoritmo de extracción de características dentro de las cuales destacan por su uso común la técnica "Log-mel Filter Bank Energies"(LFBE) o "Mel-frequency Cepstral Coefficients"(MFCC), que alimenta un clasificador el cual produce probabilidades para las n salidas [5]. De igual manera, los clasificadores tradicionales utilizados en estos sistemas de RCH son los Modelos de Cadenas de Markov Ocultas (MCMO) para segmentos de audio clave y no-clave y los algoritmos de Viterbi para la búsqueda de mejor ruta en el grafo de decodificación [4]. Aunque esta técnica aún ofrece resultados atractivos, puede ser costosa computacionalmente dependiendo de la topología. Consecuentemente, algunos estudios han reemplazado el uso de MCMO por el uso de RNA recurrentes (RNAR) [1] alcanzado mejores resultados en términos de exactitud pero con un desempeño no adecuado en términos de latencia de detección.

Los sistemas que utilizan video detección usualmente utilizan una serie completa de datos sin tomar en cuenta que éstos contienen información redundante la mayoría del tiempo con el consecuente uso sub-óptimo de los recursos computacionales. Algunas propuestas sugieren el uso de imagenes clave o representativas de la secuencia de video para reducir la cantidad de procesamiento al seleccionar las imagenes clave basados en la discriminación potencial por entropía [6]. Para habilitar los sistemas siempre alertas con un consumo energético bajo, algunas propuestas sugieren utilizar técticas de Compresión de Dominio (CD) la cual permite la ejecución de las rutinas de detección con diferentes imagenes en dos capas; la primera para reducir la resolución de la imagen y la segunda para transferir ésta a el dominio comprimido. Enseguida, es posible calcular el centro de movimiento y ejecutar un clasificador de tipo vecino más cercano ("nearest neighbor").

3. Método

El método propuesto en este trabajo consiste en un algoritmo que permite activar (o "des-bloquear") sistemas de cómputo (por ejemplo prendas electrónicas de realidad virtual) ulitizando video detección de gestos manuales de manera eficiente. Cuando el sistema está bloqueado, por ejemplo, en modo de ahorro de energía, el algoritmo propuesto monitorea cada cuadro de imagen de video capturado para determinar si el usuario realiza un gesto manual de activación o desbloqueo. Los pasos en general que sigue el algoritmo se muestran en el diagrama de alto nivel de la figura 1 y se listan a continuación:

1. El algoritmo captura la imágen F, de tamaño $n \times m$ y calcular la métrica de color promedio U llamada "tono de piel". Esta métrica se calcula dentro de una pequeña sección W de la imágen F llamada "ventana de detección"
2. La posición de W se predefine dependiendo del tipo de aplicación (por ejemplo, al centro de la imágen F en el caso de prendas electrónicas de realidad virtual como HoloLens)
3. En seguida, el algoritmo utiliza U para generar una imagen G de tamaño $R \times P$ (más pequeña y simple que la imagen original) para ser evaluada. Esta imagen está formada por el conjunto de todos los pixeles conectados por un color similar a U, comenzando dentro de la "ventana de detección"W y finalizado cuando la conexión de color se pierde
4. para evitar los problemas generados por las diferentes condiciones de iluminación, G es transformada en una imágen binaria. Consecuentemente, G contiene un único objeto con una única tonalidad
5. Adicionalmente, para prevenir problemas de escala, G es ajustada a un tamaño final resultando en una imagen G', de tamaño $M \times M$ reducido y predefinido dependiendo de la aplicación para su evaluación para determinar la existencia de un gesto válido.
6. La evaluación puede ser realizada por una RNA o cualquier otro algoritmo de decisión/clasificación. El sistema es activado o desbloqueado solo si G' con tiene un gesto válido.

El costo computacional es mínimo comparado con el estado del arte debido a un proceso simplificado de segmentación de tono de piel. Esto es, se obtiene una pequeña imagen binaria por cada imagen capturada de video lo cual reduce significativamente el costo computacional y también mejora el desempeño de detección. La eficiencia del algoritmo propueso se basa en el uso del proceso dinamico de segmentación de piel descrito graficamente en la Fig. 2. El algoritmo utiliza el espacio de color Tint, Saturation y Lightness (TSL) ya que permite una mejor segmentación del tono de piel. En contraste con los métodos tradicionales, la propuesta realiza la conversión de color solo para un numero reducido de pixeles, no para la imagen completa. Esto significa que nuestro algoritmo tomo una pequeña ventana, por ejemplo, centrada para calcular el promedio y la distribución de color. Estos valores se requieren para calcular la distancia de Mahalanobis dada por la distancia de una observación $x = (x_1, x_2, x_3 \cdot x_N)$, dentro de un conjunto de observaciones con media $u = (u_1, u_2, u_3 \cdot u_N)$ y matriz de covarianza S definida como $d(x) = \sqrt{(}((x-u)S^{(}-1)(x-u))$. Comenzando desde el centro, se aplica el algoritmo de rellenado por desborde ("flood fill") para convertir a TSL solo los pixeles vecinos del pixel original. Solo los pixeles con un color similar y conectados con los pixeles previos se incluyen en la máscara de segmentación. Al final de este proceso, se genera un sólo objeto con todos los pixeles adyacentes, conectados y de color similar. De esta manera, el algoritmo entrega un solo patrón por imágen para su subsequente evaluación.

Nuestro método utiliza la pila (o estack) (ver Figura 3) para agregar los píxeles adyacentes al píxel bajo análisis solo si tienen un color similar. Terminando

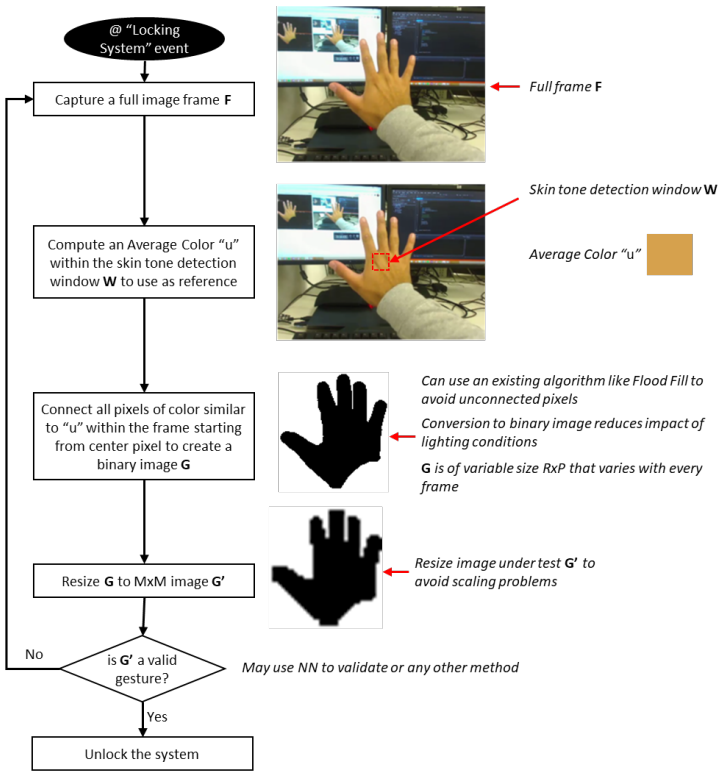

Fig. 1. El algoritmo propuesto permite una video detección de gestos manuales de bajo coste energético para la activación o desbloqueo de sistemas de cómputo

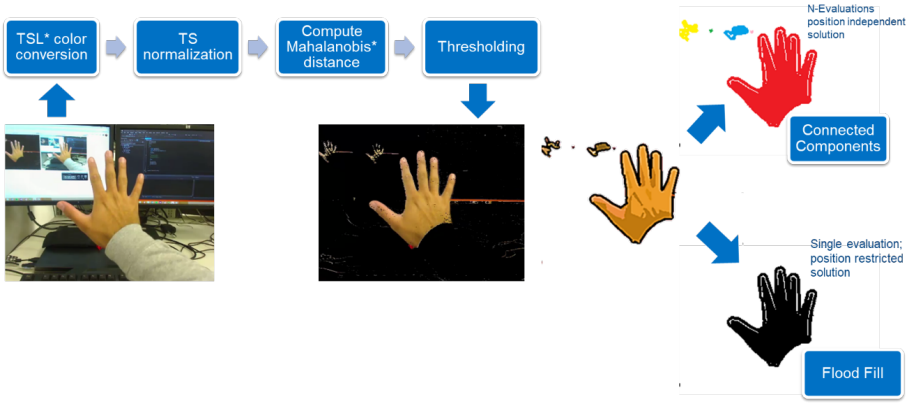

Fig. 2. Diagrama de bloques que muestra cómo se realiza la segmentación del tono de piel dinámico sobre el centro de la imagen.

con un objeto único con todos los componentes conectados, para ser analizados por una NN o cualquier otro tipo de clasificador de aprendizaje automático.

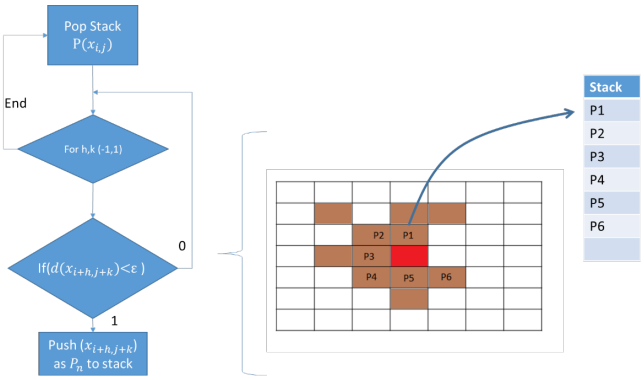

Fig. 3. Stack de píxeles basados en color similar. Esta metodología común se utiliza en nuestro método de segmentación de la mano.

4. Resultados

Se observó experimentalmente que el uso de Stack en nuestro método reduce el procesamiento de 28 evaluaciones por cuadro a solo una evaluación por cuadro. Nuestro método permite redimensionar dinámicamente el objeto segmentado (por ejemplo, la mano del usuario) a un patrón de imagen normalizado. Permitiendo la multiescala. Además, como el color se adquiere en cada imagen, cada imagen tiene en cuenta las condiciones de luz y el color de la piel.Esto significa que el algoritmo de desbloqueo es lo suficientemente robusto como para no verse afectado si el usuario usa guantes de diferente color que el tono de piel. Para mejorar y facilitar la experiencia del usuario, el sistema utiliza la realidad aumentada para mostrar el patrón de "despertar"de desbloqueo en la imagen donde el usuario tiene que alinear el gesto de la mano, como se muestra en la Figura 4.

Se implementó la prueba de concepto de algoritmo de desbloqueo descrito anteriormente. Es importante tener en cuenta que la NN que hemos experimentado reconoce un solo patrón. Una vez que se detecta este patrón de gesto de la mano, el sistema se activa, en este punto, el sistema puede cambiar a un modelo de reconocimiento NN más grande para reconocer un mayor número de gestos con las manos según la aplicación, detectar cualquier posición en la pantalla e incluso reconocer simultáneamente las dos manos de los usuarios.

Este método también comprende una subrutina para reducir aún más el costo computacional por medios estadísticos. Esta reducción se logra ya que en algunos otros casos, el algoritmo de desbloqueo también está considerando evitar

Fig. 4. Gesto de la palma abierta utilizado por el algoritmo de desbloqueo para activar el sistema, presentando un patrón de gestos como realidad aumentada para dar al usuario una indicación del uso.

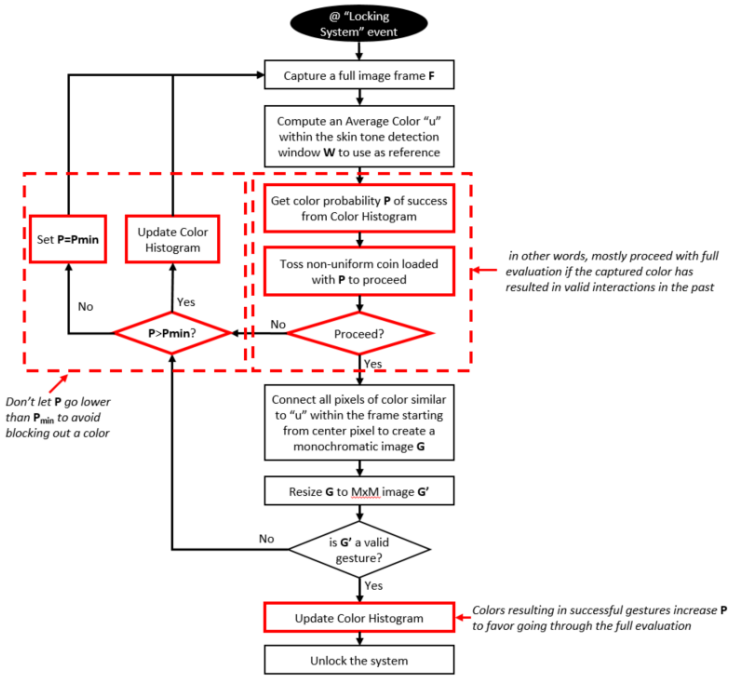

Fig. 5. El algoritmo también puede evitar la necesidad de evaluar cada cuadro utilizando el conocimiento estadístico de la probabilidad de tono de piel para dar como resultado una detección exitosa del gesto de la mano. Las secciones rojas de este diagrama resaltan las modificaciones al algoritmo de la figura 1 para lograr una reducción en el costo computacional.

la necesidad de realizar el proceso de tono dinámico para cada fotograma.En cambio, una vez que se obtiene un tono de piel, el método utiliza un modelo estadístico para determinar si procede con la segmentación dinámica del tono de piel en función de la probabilidad de tener un tono de la segmentación válida, es decir, se realiza la segmentación del tono de piel en colores que resultaron en interacciones válidas en el pasado. La Figura 5 muestra una versión actualizada en el algoritmo de desbloqueo descrito en la Figura 1, donde después de capturar el tono de piel, se actualiza una distribución estadística de colores. Si se captura un nuevo tono de piel, la probabilidad de su uso comienza en 50/50, y aumenta su probabilidad con el tiempo en función de su uso válido, o disminuye cuando se produce una interacción invalida.

5. Conclusiones

Este método proporciona un algoritmo de desbloqueo computacionalmente eficiente para (despertar) activación de sistemas basados en visión por computadora mediante la detección de un comando de gesto con la mano. El método que proponemos extrae un solo objeto por cuadro para la evaluación usando un clasificador, por ejemplo, haciendo segmentación de color de piel dinámica. Este método evita la necesidad de evaluar cada cuadro de video mediante el desarrollo de juicios estadísticos de detección de mano exitosa. Nuestro método produce una reducción significativa en términos de recursos computacionales cuando se compara con los métodos tradicionales de DL, Además muestra robustez a diferentes condiciones de iluminación y diferentes colores de tono de piel (incluido el uso de guantes), y funciona para diferentes tamaños o distancias de la mano del usuario (escala múltiple dinámica).

Referencias

1. Fernández, S., Graves, A., Schmidhuber, J.: An application of recurrent neural networks to discriminative keyword spotting. In: de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D. (eds.) Artificial Neural Networks – ICANN 2007. pp. 220–229. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
2. Köpüklü, O., Gunduz, A., Kose, N., Rigoll, G.: Real-time hand gesture detection and classification using convolutional neural networks. CoRR abs/1901.10323 (2019), http://arxiv.org/abs/1901.10323
3. Maghoumi, M., Jr., J.J.L.: Deepgru: Deep gesture recognition utility. CoRR abs/1810.12514 (2018), http://arxiv.org/abs/1810.12514
4. Rohlicek, J.R., Russell, W., Roukos, S., Gish, H.: Continuous hidden markov modeling for speaker-independent word spotting. In: International Conference on Acoustics, Speech, and Signal Processing,. pp. 627–630 vol.1 (May 1989)
5. Zhang, Y., Suda, N., Lai, L., Chandra, V.: Hello edge: Keyword spotting on microcontrollers. CoRR abs/1711.07128 (2017), http://arxiv.org/abs/1711.07128
6. Zhao, Z., Elgammal, A.M.: Information theoretic key frame selection for action recognition. In: BMVC (2008)