


Reconocimiento de acciones humanas en tiempo real en entornos abiertos

Mayra Vanessa Alvear Gallón 

maalvear@unirioja.es

Dpto. de Matemáticas y Computación, Universidad de La Rioja

Palabras clave: Gesture recognition, uncontrolled environment, contactless interface, computer vision

El lenguaje gestual y corporal supone un mecanismo de comunicación fundamental entre seres humanos. Tratar que las máquinas entiendan este lenguaje es un área activa de investigación en el campo de la visión por computación que tiene su aplicación en campos diversos como la videovigilancia o la sanidad. Sin embargo, la mayoría de estos desarrollos se realizan en entornos cerrados y controlados. El reconocimiento de movimientos y gestos en tiempo real a través de cámaras en un entorno no controlado (como el vestíbulo de una universidad o la sala de un museo) y que permiten la interacción con los transeúntes en estos espacios presenta retos en diversos ámbitos que incluyen, al menos, desafíos tecnológicos, sociales y legales que deben abordarse cuidadosamente. El proyecto de tesis descrito se enmarca en esta línea de investigación. Consiste en un Doctorado Industrial fruto de la colaboración entre la empresa IR Soluciones y la Universidad de La Rioja. La hipótesis de partida consiste en que es posible el desarrollo de modelos de inteligencia artificial que permitan el reconocimiento gestual en entornos abiertos a través de una interfaz sin contacto. Esta hipótesis se explora a través de dos objetivos estrechamente relacionados. El primero es el diseño de una interfaz web adaptada a la interacción sin contacto físico mediante imágenes de vídeo capturadas con una cámara en tiempo real. El segundo persigue la construcción de modelos de inteligencia artificial que garanticen, en un entorno abierto, una interacción con dicha interfaz.

Como primer paso, se estudiaron casos de uso de interfaces sin contacto que permitan realizar una interacción humano-máquina. El desarrollo de estos ejemplos se basan en un estudio de las técnicas contenidas en la li-

Este trabajo está parcialmente financiado por un Doctorado industrial, Comunidad Autónoma de La Rioja, y por el proyecto PID2020-115225RB-I00 financiado por MCIN/AEI/ 10.13039/501100011033.

teratura, la creación de un dataset que nos permita entrenar modelos de aprendizaje profundo y finalmente, la integración de dichos modelos en un sistema de reconocimiento de diferentes signos.

Este proyecto está pensando para que tenga un impacto relevante tanto en la propia empresa donde se desarrolla como en la investigación realizada desde la universidad, sin perder de vista que esta tecnología puede impactar finalmente en la sociedad.

Introducción

La visión por computación es una rama de la inteligencia artificial que tiene como objetivo dotar a los ordenadores de la capacidad de obtener información a partir de imágenes y vídeos. En los últimos años, se ha producido una gran revolución en este campo debido al uso de las técnicas de aprendizaje profundo. Un área de investigación activa dentro de la visión por computación es el reconocimiento de gestos y acciones humanas, de esta manera, en la última década se ha llevado a cabo un amplio trabajo de investigación, como puede verse en las literaturas existentes [1], [2], [3], [4]. Incluso, la reciente difusión de sistemas de cámaras de vídeo de bajo costo, incluidas las cámaras de profundidad [5], ha reforzado el desarrollo de sistemas que requieren del reconocimiento de acciones humanas en tiempo real necesarias en diversas aplicaciones como la videovigilancia, la seguridad doméstica, la sanidad, entre otras [1]. Sin embargo, la mayoría de estos desarrollos se llevan a cabo en entornos cerrados y controlados.

El reconocimiento de movimientos y gestos en tiempo real a través de cámaras que adquieren sus imágenes en un entorno no controlado (como un centro comercial, el vestíbulo de una universidad, el vestíbulo de un museo, el andén de una estación de tren o la terminal de un aeropuerto) y que permite la interacción con los transeúntes en estos espacios, presenta retos en diversas áreas que incluyen, al menos, desafíos tecnológicos, sociales y legales que deben abordarse cuidadosamente.

En el ámbito tecnológico, es necesario desarrollar modelos que puedan ser utilizados en diferentes espacios no controlados. Así mismo, se pueden utilizar diferentes técnicas como el aumento de datos [6], o la transferencia de modelos [7], las cuales pueden complementar la adquisición y anotación de conjuntos de datos suficientemente generales para poder entrenar y validar estos modelos. Más aún, dentro de estos aspectos técnicos, existen dificultades derivadas del ruido en un espacio no controlado (personas que pasan por detrás de la que está delante del foco de la cámara). Por tal motivo, las cámaras deben tener en cuenta los diferentes niveles de profundidad para poder enfocar eficazmente el aspecto que se quiere estudiar [5].

En cuanto a los factores sociales, la interfaz debe diseñarse cuidadosa-

mente cuando se generen los movimientos o gestos ante la cámara, para que se interpreten como el usuario espera para evitar sentimientos de frustración (o incluso rechazo) que podrían generarse si los movimientos o gestos ante la cámara no se interpretan como el usuario espera. Por último, se debe tener en cuenta los aspectos legales para analizar cómo deben instalarse estas cámaras en espacios no controlados y la solicitud de los permisos necesarios para los usuarios en el momento de captar las imágenes.

En este trabajo se describe un proyecto de tesis que se enmarca en esta línea de investigación relacionada con el reconocimiento de acciones humanas en tiempo real en entornos abiertos. Consiste en un Doctorado Industrial fruto de la colaboración entre la empresa Innovación Riojana de Soluciones S.L.U y la Universidad de La Rioja.

La hipótesis de partida consiste en que es posible el desarrollo de modelos de inteligencia artificial que permitan el reconocimiento gestual en entornos abiertos a través de una interfaz sin contacto. Creemos que la investigación que exige el estudio de esta hipótesis supone retos que pueden abrir líneas de negocio en la empresa y que permitirán transferir a la sociedad los resultados obtenidos a través de la misma.

Esta hipótesis la pretendemos explorar a través de los siguientes dos objetivos generales del proyecto, que están estrechamente relacionados:

- OG1. Diseño de una interfaz web adaptada a la interacción sin contacto físico (es decir, a través de imágenes de vídeo capturadas con una cámara en tiempo real).
- OG2. Construcción de modelos de inteligencia artificial (basados en *deep learning*) que garanticen, en un entorno abierto, una interacción con la interfaz cuyo diseño se recoge en el primer objetivo general.

Estos objetivos generales presentan importantes desafíos. Por una parte, parece poco realista abordar el diseño de interfaces generales, por lo que planeamos centrarnos en alguna concreta que sea de interés para el ámbito de negocio de la empresa. En segundo lugar, la pretensión de que el sistema funcione en un entorno abierto (es decir, que las cámaras adquieran sus imágenes en circunstancias en las que un flujo de personas pasa ante ellas) es demasiado ambiciosa.

Trabajos preliminares

Con este fin, se han realizado los siguientes estudios:

1. Estado del arte: se ha realizado una revisión de literatura sobre reconocimiento de acciones y de gestos.

2. Estudio detallado de modelos de Transformers aplicados a la visión por computación.
3. Estudio de cámaras, bases de datos y herramientas actuales en el mercado.
4. Entrenamiento de modelos de clasificación utilizando diferentes librerías de aprendizaje profundo.
5. Estudio de herramientas para crear una base de datos.

Adicionalmente, para explorar la literatura revisada se han implementado algunos casos de uso que se ahondarán más en la charla:

- Interacción sin contacto para ordenar una hamburguesa
- Hand Tracking interaction 3D usando Unity Hub
- Reconocimiento en tiempo real de acciones de la Lengua de Signos Española (LSE) usando LSTM en un entorno controlado.
- Reconocimiento en tiempo real de las vocales del LSE en un entorno controlado.
- Uso del mouse sin contacto a través de 4 gestos de la mano enfocado para interfaces en vertical.

Bibliografía

- [1] Tansel Özyer, Duygu Selin Ak, and Reda Alhajj. Human action recognition approaches with video datasets—A survey. *Knowledge-Based Systems*, 222:106995, 2021.
- [2] Upal Mahbub and Md Atiqur Rahman Ahad. Advances in human action, activity and gesture recognition. *Pattern Recognition Letters*, 155:186–190, 2022.
- [3] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, 2019.
- [4] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, 32:200901, 2020.

- [5] Arya Sarkar, Avinandan Banerjee, Pawan Kumar Singh, and Ram Sarkar. 3D human action recognition: Through the eyes of researchers. *Expert Systems with Applications*, page 116424, 2022.
- [6] Ángela Casado-García, César Domínguez, Manuel García-Domínguez, Jónathan Heras, Adrián Inés, Eloy Mata, and Vico Pascual. CLoDSA: a tool for augmentation in classification, localization, detection, semantic segmentation and instance segmentation tasks. *BMC Bioinformatics*, 20(1):1–14, 2019.
- [7] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.