Deep Learning para detección y tracking de objetos en vídeo

Brais Bosquet, Mauro Fernández-Sanjurjo, Lorenzo Vaquero, Daniel Cores, Víctor Brea, Manuel Mucientes
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS). Universidade de Santiago de Compostela.
manuel.mucientes@usc.es

La detección de objetos mediante visión por computador requiere enmarcar el conjunto de objetos de interés que aparecen en una imagen y clasificarlos en alguna de las categorías predefinidas. Por otra parte, el tracking de objetos trata de mantener la identidad de cada uno de ellos a lo largo de un vídeo. En este trabajo se aborda la realización de ambas tareas mediante redes convolucionales generadas de forma automática mediante técnicas de aprendizaje profundo. Concretamente nos centraremos en tres líneas: (i) la detección de objetos pequeños (menos de 16x16 píxeles) mediante la red STDnet [1]; (ii) el tracking de objetos mediante una red convolucional que opera en tiempo real y con múltiples objetos; y (iii) la monitorización de tráfico mediante un sistema integral de detección y tracking.

La red STDnet (Small Target Detection network) incluye un mecanismo de atención visual temprana, denominado RCN (Region Context Network), que permite seleccionar las regiones más prometedoras que contienen objetos pequeños y su contexto. La RCN permite trabajar con mapas de características con alta resolución, pero utilizando una cantidad de memoria reducida. Los mapas de características filtrados, que contienen las regiones con mayor probabilidad de contener objetos, son procesados a lo largo de la red hasta finalizar en una RPN (Region Proposal Network) previa a la clasificación final. La RCN es clave para (i) incrementar la precisión en la localización de los objetos ---al permitir trabajar con mapas de características de alta resolución---, (ii) reducir el tamaño de la red en memoria, y (iii) aumentar la velocidad de procesado de la red.

Por otra parte, el sistema integral de detección y tracking ha sido desarrollado para operar en tiempo real con centenares de objetos y ser robusto a oclusiones totales. Para ello, el sistema realiza el tracking mediante dos componentes: (i) el tracking de bajo nivel, basado únicamente en características visuales del objeto; (ii) el tracking de alto nivel, que permite tener en cuenta el modelo de movimiento del objeto, realizar la asociación de datos, y gestionar la reinicialización del tracker de bajo nivel.

Elementos clave:

- STDnet, una red convolucional para detectar objetos de menos de 16x16 píxeles.
- Una red convolucional para el tracking de múltiples objetos en tiempo real.
- Un sistema integral de detección y tracking robusto a oclusiones.

Referencias

[1] B. Bosquet, M. Mucientes, and V. Brea. STDnet: A ConvNet for Small Target Detection. In Proceedings of the 29th British Machine Vision Conference (BMVC), Newcastle (UK), 2018.