Tesis de Maestría

Reconocimiento de acciones humanas en video, utilizando detección de contornos y Deep Learning

Cronograma de Trabajo

Alumno: Rodrigo Fondato

Tutor: Juliana Gambini

1. Descripción del problema

El reconocimiento de actividades humanas (HAR, por sus siglas en inglés), es la habilidad para interpretar movimientos y gestos efectuados por personas a través de sensores, para luego determinar qué acciones o actividades están realizando (Ann and Theng, 2014).

Está técnica tiene múltiples aplicaciones, como la predicción de actividades peligrosas o delictivas a partir de secuencias de video tomadas por cámaras instaladas en la vía pública (Ryoo, 2011), aplicaciones médicas, como un sistema para el monitoreo de ejercicios de rehabilitación (González-Ortega et al., 2014) o la generación de instrucciones para interactuar con un software mediante el reconocimiento de gestos y movimientos de los usuarios (Gerling et al., 2012).

A partir del desarrollo del hardware de aceleración gráfica (GPUs) y la publicación de AlexNet (Krizhevsky et al., 2012), las redes profundas tomaron protagonismo en el campo de visión por computadora. Muchas investigaciones recientes en el campo del reconocimiento de acciones humanas están mayormente basadas en el uso de redes convolucionales, debido a que permiten extraer descriptores robustos en forma automática a partir de los datos de entrenamiento (Ji et al., 2013; Feichtenhofer et al., 2016; Varol et al., 2018; Ullah et al., 2017; Zhu et al., 2019).

Sin embargo, requieren una gran cantidad de recursos computacionales (poder de procesamiento y memoria), tiempo y disponibilidad de datos para su entrenamiento, debido a la cantidad de parámetros que deben ser ajustados (Li et al., 2016).

Si bien una ventaja de las redes convolucionales es que la cantidad de parámetros no depende del tamaño de los datos de entrada (ya que los pesos se comparten), algunas investigaciones demuestran que sí existe una correlación entre dicho tamaño y la complejidad óptima de la red, necesitando una mayor cantidad de capas o filtros al incrementar la resolución de entrada. Esto puede explicarse dado que, a una mayor resolución de imagen, son necesarias más capas para obtener campos receptivos más amplios que permitan capturar características que consideren una cantidad superior de píxeles (Tan and Le, 2019).

Para el caso particular del procesamiento de video, a las dos dimensiones temporales de entrada (alto y ancho) se le suma una tercera dimensión, el tiempo, medido en cantidad de cuadros. Para procesar entradas de datos de ésta dimensionalidad, fueron propuestas redes convolucionales que utilizan filtros 3D, denominadas C3D. Debido a la dimensión adicional en los filtros, la cantidad de parámetros de éstas redes pueden escalar exponencialmente, demandando un mayor poder de cómputo (Li et al., 2019).

Para evitar el aumento de la complejidad de la red, es necesario conservar un determinado tamaño en los datos de entrada. Por éste motivo, existe un compromiso entre la resolución de las imágenes de entrada y la cantidad de cuadros consecutivos que serán considerados para la detección de acciones. A modo de ejemplo, uno de los modelos propuestos utiliza solo 9 cuadros consecutivos para efectuar la clasificación (Ji et al., 2013). Un método para aumentar el horizonte temporal es disminuir la resolución de los videos utilizados (Asadi-Aghbolaghi et al., 2017).

En muchas aplicaciones, la detección debe ser realizada en tiempo real. Para éstos casos hay un restricción en el tiempo total disponible para realizar el procesamiento.

El problema que se intenta resolver es entonces la detección computacionalmente eficiente de acciones humanas básicas en video, mediante el uso de contornos humanos como datos de entrada, permitiendo la reducción de la cantidad de parámetros a ajustar y consecuente complejidad del modelo.

2. Objetivos específicos

Se propone realizar las siguientes actividades:

1. Definición del proceso de gestión del proyecto, selección de herramientas a utilizar para la gestión y creación de lista de tareas.
2. Definición del proceso de revisión de la literatura y fuentes a investigar, armado de consultas y criterios de selección de artículos, y ejecución del proceso de selección. Lectura de los artículos, y armado del estado de la cuestión.
3. Identificación, clasificación, selección y obtención de bases de datos públicas que contengan segmentos de video que muestren acciones humanas.
4. Análisis y selección de los segmentos de video según categorías, calidad, duración y formato.
5. Implementación y ejecución de un proceso para adaptar y homogeneizar los distintos archivos, obteniendo como resultado una única base de datos.
6. Diseño e implementación del modelo de reconocimiento de acciones humanas.
7. Diseño de las métricas e implementación de la herramienta que las obtendrá.
8. Entrenamiento y optimización del modelo.
9. Revisión del diseño de las características y el modelo, propuesta e implementación de mejoras. Obtención de nuevas métricas y elección del modelo ganador.
10. Instalación y ejecución de modelos pre-entrenados existentes. Obtención de métricas.
11. Análisis de los resultados obtenidos en la comparación de modelos. Elaboración de conclusiones y finalización del documento de tesis.

3. Cronograma de trabajo

El Cronograma de Trabajo se muestra en el siguiente cuadro:

Actividad	Mes										
	1	2	3	4	5	6	7	8	9	10	11
Actividad 1	x										
Actividad 2	x	x	x								
Actividad 3				x							
Actividad 4				x	x						
Actividad 5					x	x					
Actividad 6						x	x	x			
Actividad 7								x			
Actividad 8									x		
Actividad 9									x	x	
Actividad 10										x	x
Actividad 11											x

4. Bilbiografía

Ann, O. C. and Theng, L. B. (2014). Human activity recognition: A review. In *Proceedings - 4th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2014*, pages 389–393.

Asadi-Aghbolaghi, M., Clapes, A., Bellantonio, M., Escalante, H. J., Ponce-Lopez, V., Baro, X., Guyon, I., Kasaei, S., and Escalera, S. (2017). A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences. In *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*, pages 476–483.

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016). Convolutional Two-Stream Network Fusion for Video Action Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 1933–1941.

Gerling, K. M., Livingston, I. J., Nacke, L. E., and Mandryk, R. L. (2012). Full-body motion-based game interaction for older adults. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 1873–1882.

González-Ortega, D., Díaz-Pernas, F. J., Martínez-Zarzuela, M., and Antón-Rodríguez, M. (2014). A Kinect-based system for cognitive rehabilitation exercises monitoring. *Computer Methods and Programs in Biomedicine*, 113(2):620–631.

Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3D Convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, J., Jiang, X., Sun, T., and Xu, K. (2019). Efficient violence detection using 3D convolutional neural networks. In *2019 16th IEEE International Confe-*

rence on Advanced Video and Signal Based Surveillance, AVSS 2019. Institute of Electrical and Electronics Engineers Inc.

Li, X., Zhang, G., Huang, H. H., Wang, Z., and Zheng, W. (2016). Performance analysis of gpu-based convolutional neural networks. In *2016 45th International conference on parallel processing (ICPP)*, pages 67–76. IEEE.

Ryoo, M. S. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1036–1043.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., and Baik, S. W. (2017). Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features. *IEEE Access*, 6:1155–1166.

Varol, G., Laptev, I., and Schmid, C. (2018). Long-Term Temporal Convolutions for Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1510–1517.

Zhu, Y., Lan, Z., Newsam, S., and Hauptmann, A. (2019). Hidden Two-Stream Convolutional Networks for Action Recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11363 LNCS, pages 363–378.