Universidad Nacional Autónoma de México

Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas

Seminario de obtención de grado

Categorización semántica de objetos a partir de la Interacción Humano-Objeto.

Edgar de Jesús Vázquez Silva

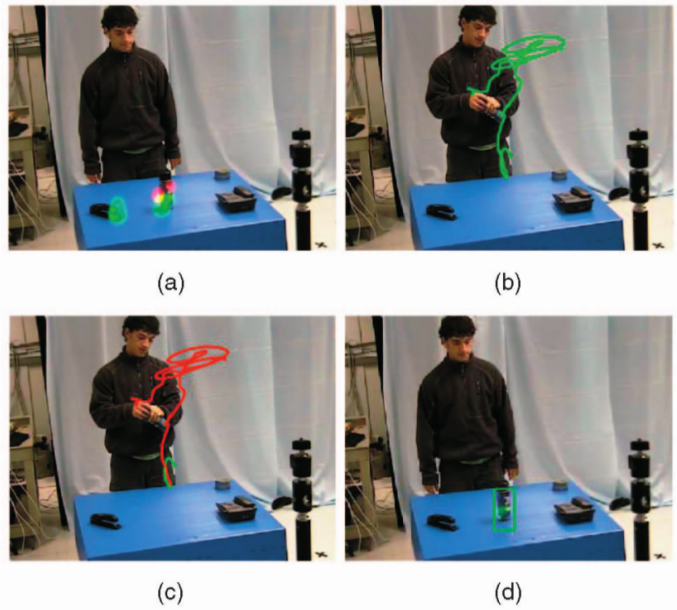
Tutor: Jesus Savage Carmona

11 de agosto de 2020

Contenido

Definición del problema

Motivación



Interacción humano objeto. (a) Segmentación de objetos similares a una botella, (b) Trayectoria de manipulación de objetos, (c) Trayectoria de interacción con objeto, verde el agarre del objeto, rojo la manipulación de uso. (d) Categorización del objeto botella despúes de la interacción.

Definición del problema

Definición del problema

- Se pretende tener un sistema capaz de inferir las propiedades de uso de un objeto a partir de la información proveniente de cámaras RGB.
- El sistema debe ser utilizado en un robot de servicio doméstico, haciendo uso de cámaras RGB.
- Los objetos pueden presentar caracteristicas visuales similares, en tal caso una desambiguación es requerida.
- **Hipotesis 1**: Existe una relación entre la acción realizada por un humano y las características del objeto involucrado en tal acción.
- **Hipotesis 2**: La relación entre la acción realizada por un humano y las características del objeto involucrado en tal acción, pueden ayudar a inferir propiedades sobre los objetos.

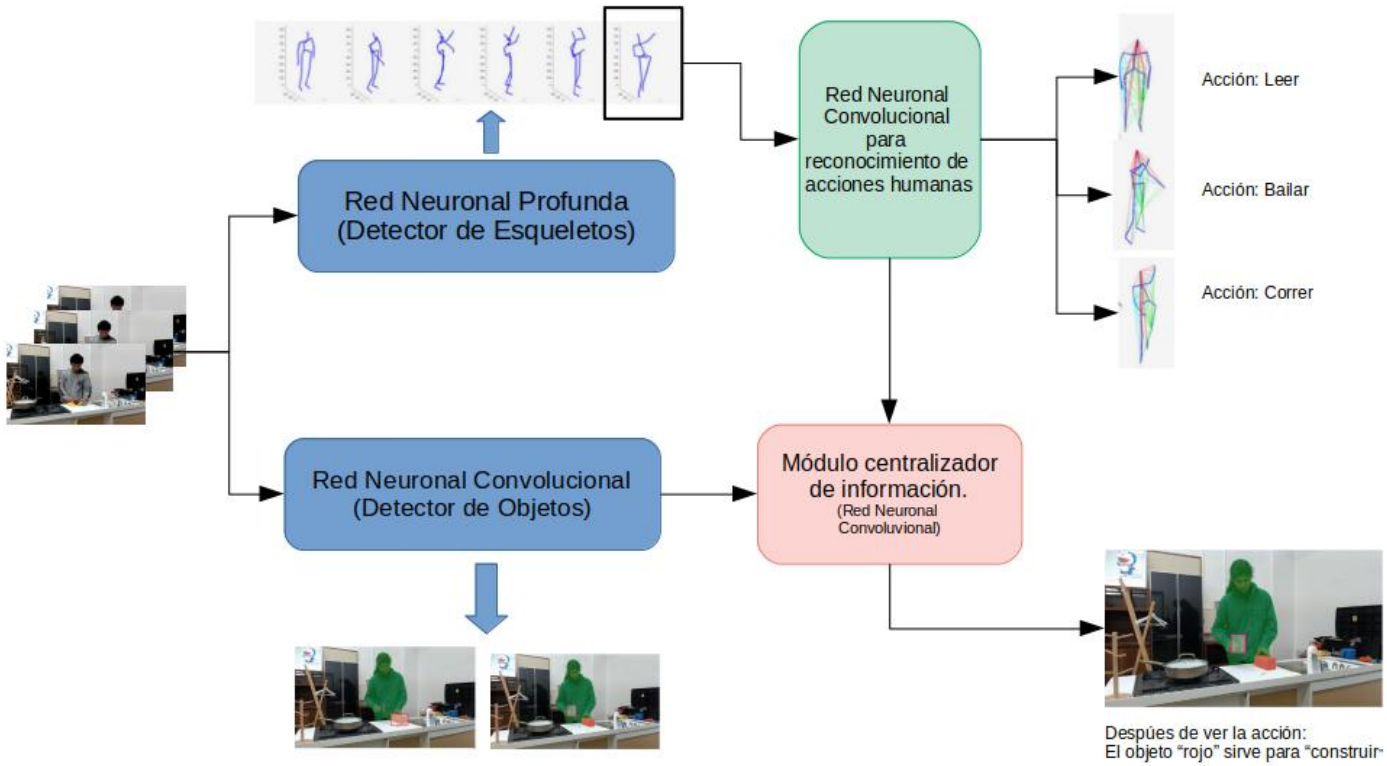
Objetivos

Objetivos

- Reconocer acciones humanas a partir de un sistema basado en redes neuronales profundas.
- Detectar objetos manipulables en una escena visual, (imagen).
- Tener un sistema que combine la información de la acciones humanas con la información de los objetos para inferir las propiedades de uso de los mismos.
- Recopilar un conjunto de datos estandarizado, debidamente etiquetado para las pruebas requeridas.

Metodología propuesta

Metodología propuesta



Descripción del sistema propuesto para la detección de interacción humano-robot.

Metodología propuesta

- Colectar un conjunto de datos en escenas de entornos domésticos de prueba que sirva como base para entrenamiento y prubas de redes neuronales propuestas.
- Como primera aproximación se abordó cada problema por separado, reconocimiento de acciones humanas y detección de objetos en la escena.
- Diseño y pruebas de una red neuronal para reconocimiento de acciones humanas, basada en esqueletos, utilizando redes neuronales convolucionales y redes neuronales recurrentes.

Metodología propuesta

- Diseño e implementación de una red neuronal convolucional para la detección de objetos.
- Entrenar y probar ambas redes en un conjuno de datos estandarizado previamente capturado.

Conjunto de datos

- El dataset se compone nubes de puntos organizadas capturadas y vídeos capturados con los siguientes dispositivos:
  - Sensor Kinect v1. (Nubes de puntos)
  - Sensor Xtion. (Nubes de puntos)
  - Camaras componentes del Robot HSR. (Nubes de puntos)
  - HD webcam C920 PRO. (Vídeos)

Conjunto de datos (Dispositivos de Captura)

(a) Logitech webcam.

(b) Sensor kinect

(c) Sensor Xtion Live Pro.

(d) Robot de servicio Doméstico HSR.

Dispositivos de captura del conjunto de datos

Conjunto de datos (Disposición de camaras)

(a)

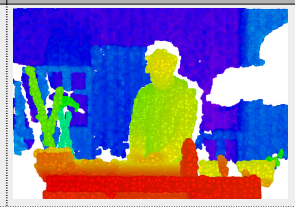
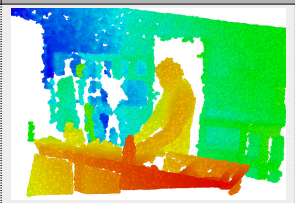
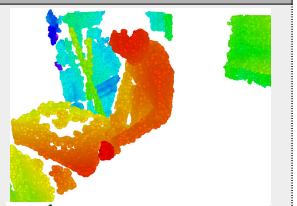
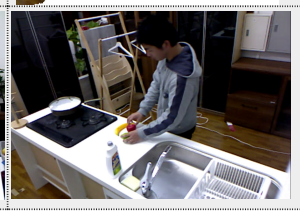
(b)

Acomodo de cámaras para captura del Dataset. Colaboración con la Universidad de Tamagawa, Tokyo, Japón durante la estancia de Investigación que comprende el periodo Enero - Marzo 2020.

Conjunto de datos

- El dataset contiene un total de 292 vídeos obtenidos desde 4 perpectivas diferentes, grabados con cada uno de los 4 dispositivos diferentes anteriormente mencionados.
- Cada uno de los 292 vídeos fue analizado y etiquetado a mano con las mascaras de los objetos y personas que aparecen en las escenas.

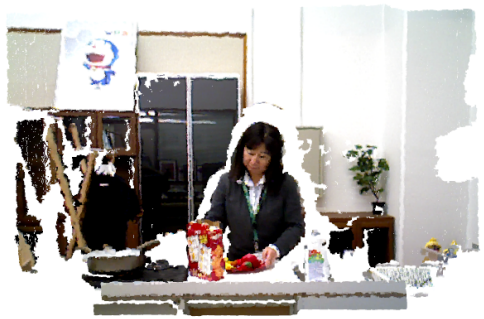
Naturaleza de la información

	Webcam	HSR	Xtion	Kinect
Depth				
Point Cloud				
Image RGB				

Ejemplo de información contenida en el Dataset. Muestra el tipo de información contenida de acuerdo al dispositivo de captura.

Conjunto de datos (Ejemplos)

(a)



(b)

(c)

Ejemplo de información contenida en el Dataset. Colaboración con la Universidad de Tamagawa, Tokyo, Japón durante la estancia de Investigación que comprende el periodo Enero - Marzo 2020.

Descripción del conjunto de datos

- Personas parcialmente visibles.
- Objetos en la escena. En particular con un dataset especifico YCB (uso en Robocup).
- Acciones.
- Interacciones con objetos.

Dataset YCB

(a)

(b)

(c)

Ejemplo de información contenida en el Dataset YCB.

Conjunto de datos (Etiquetado)

(a)

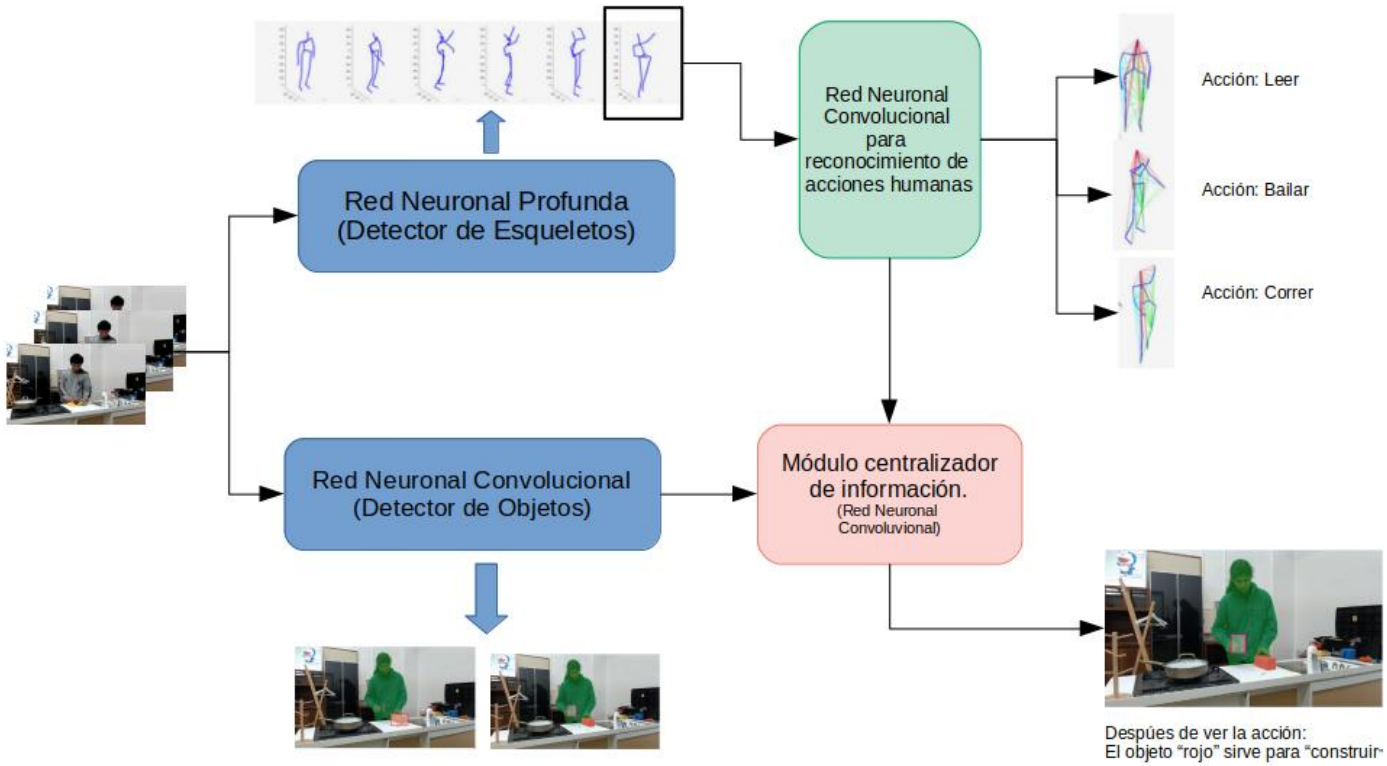
(b)

(c)

Etiquetado de objetos de interes en la escena.

Modelos

Reconocimiento de acciones



Descripción del sistema propuesto para la detección de interacción humano-robot.

CNN 2D (Reconocimiento de acciones)

- Se utilizó OpenPose como extractor detector de esqueletos.
- Se realizó *"transferencia de conocimiento"* utilizando los pesos pre-entrenados en el dataset **Kinetic**.
- En la etapa de clasificación se añadieron dos capas de neuronas totalmente conectadas, con el objetivo de hacer una clasificación multiclase.
- Por lo tanto, la última capa consta de 101 neuronas con una función de activación *softmax*.

Reconocimiento de acciones

- Extracción de cada uno de los frames componentes del vídeo.
- Finalmente se obtuvieron un total de 1,788,425 imagenes de todo el conjunto de vídeos.

Reconocimiento de acciones

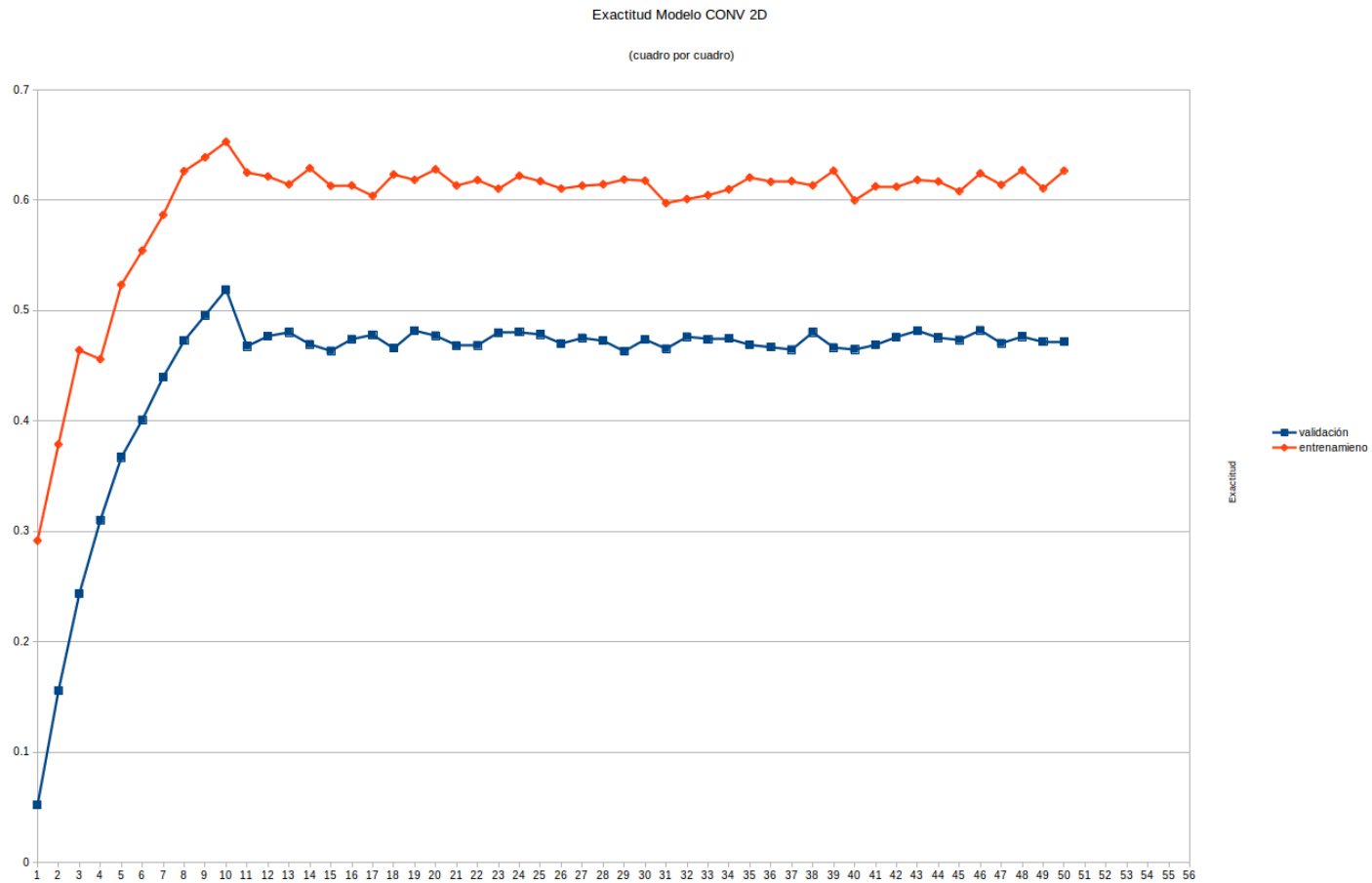
Layer (type)	Output Shape	Param #
. . .		
. . .		
==========	==========	==========
global_average_pooling2d dense_1 (Dense) dense_2 (Dense)	(None, 2048) (None, 1024) (None, 101)	0 2098176 103525

==========

Total params: 24,004,485
Trainable params: 2,201,701
Non-trainable params: 21,802,784

Resultados

Reconocimiento de acciones Conv2D

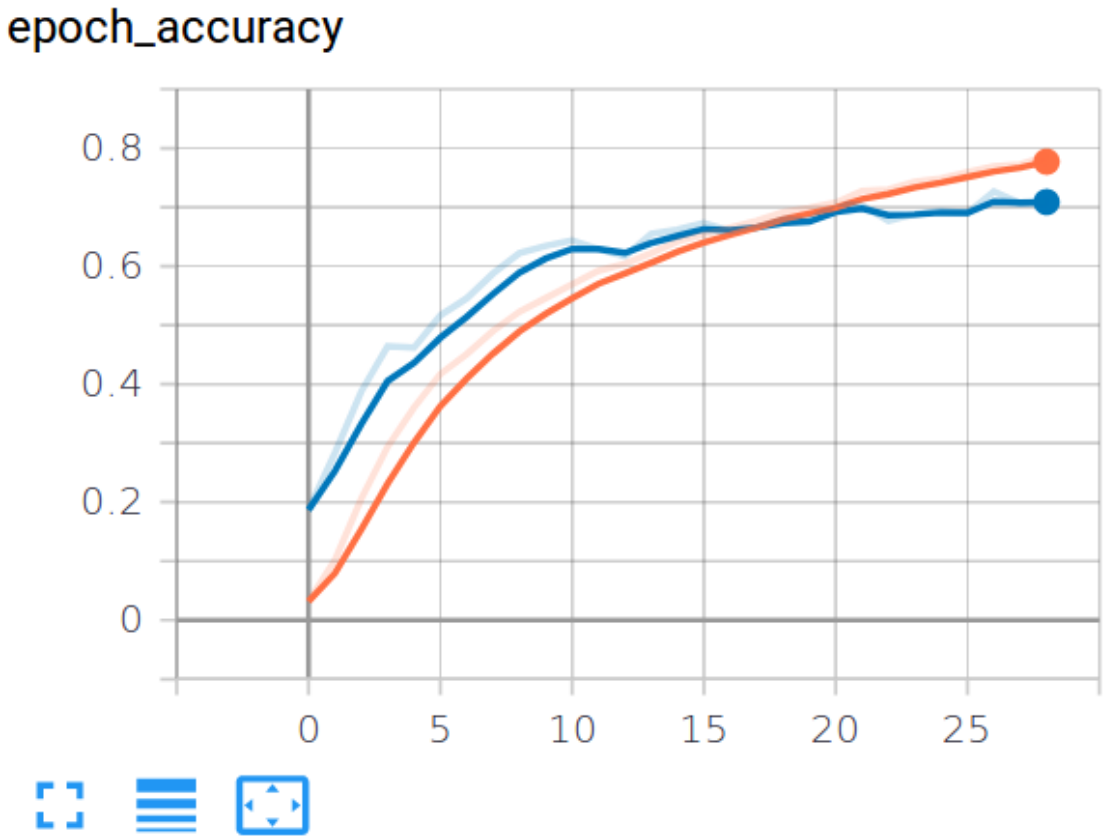


Gráfica de exactitud del modelo conv 2D cuadro por cuadro, para 51 epocas.

Reconocimiento de acciones Conv2D

- Número de parametros: 24,004,485
- Tiempo de entrenamiento: 14 horas 20 minutos (Nvidia 1050GTX)
- Porcentaje final de exactitud en validación: 0.48
- Tiempo en inferencia: 3.56 [s]

Reconocimiento de acciones RNN



Gráfica de exactitud del modelo conv 2D + LSTM, para 28 epocas.

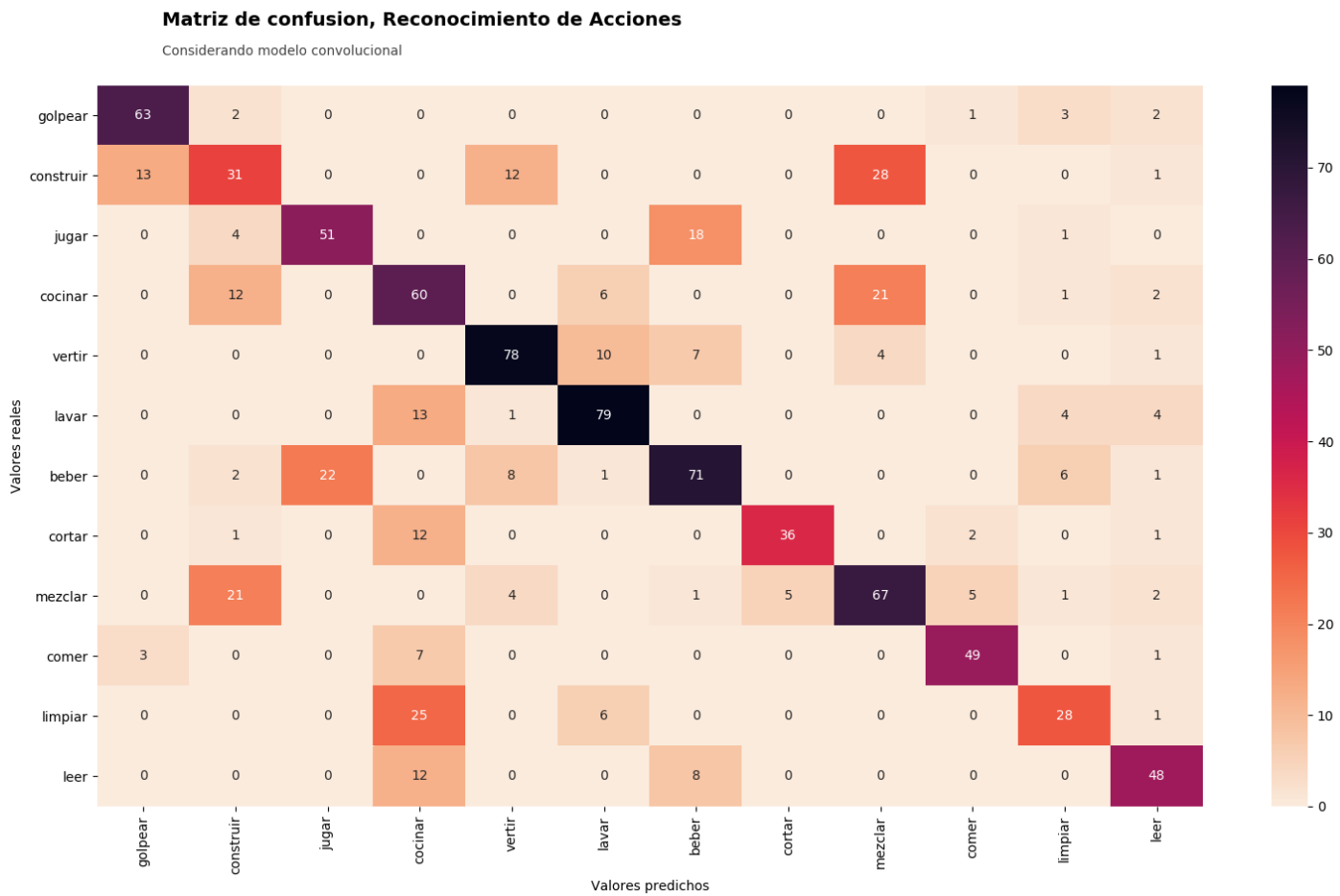
Reconocimiento de acciones RNN

Gráfica del valor de la función de perdida del modelo conv 2D + LSTM, para 28 epocas.

Reconocimiento de acciones RNN

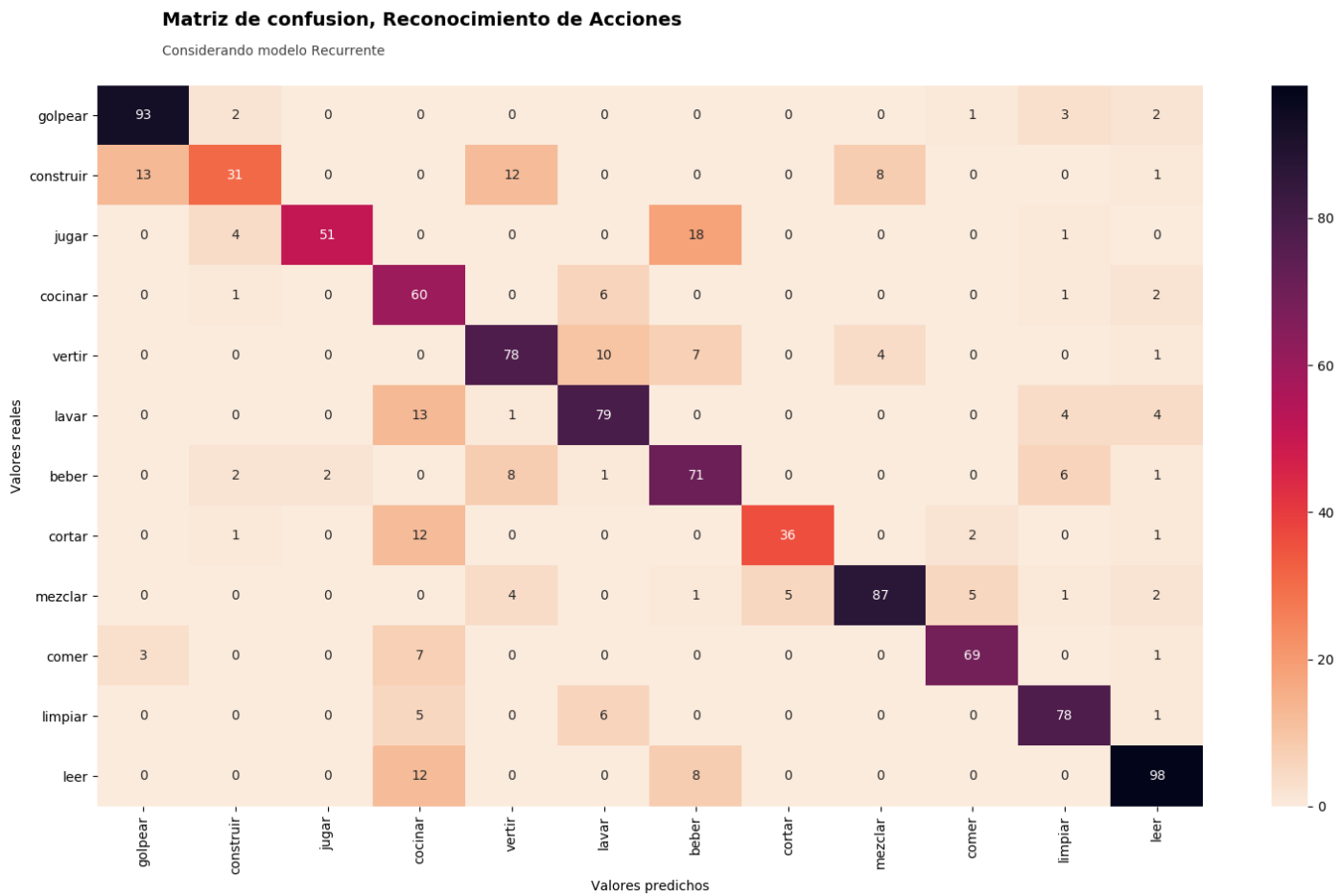
- Número de parametros: 34,663,525
- Tiempo de entrenamiento: 10 horas 34 minutos (Nvidia 1050GTX)
- Porcentaje final de exactitud en validación: 0.708
- Tiempo en inferencia: 0.8 [s]

Matriz confusión (Reconocimiento de Acciones)



Gráfica del valor de la función de perdida del modelo conv 2D + LSTM, para

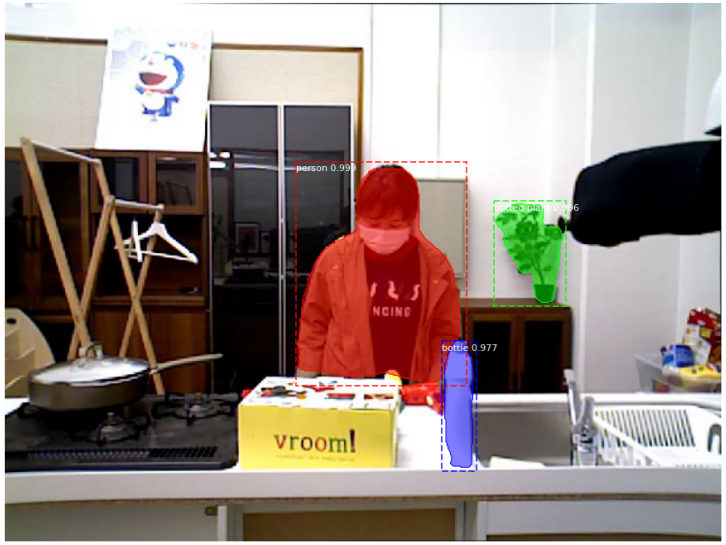
Matriz confusión (Reconocimiento de Acciones)



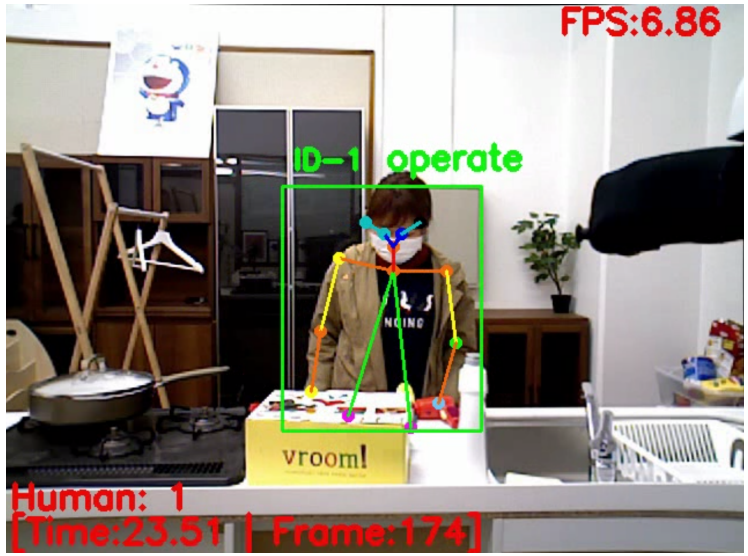
Gráfica del valor de la función de perdida del modelo conv 2D + LSTM, para

Resultados

Resultados Parciales



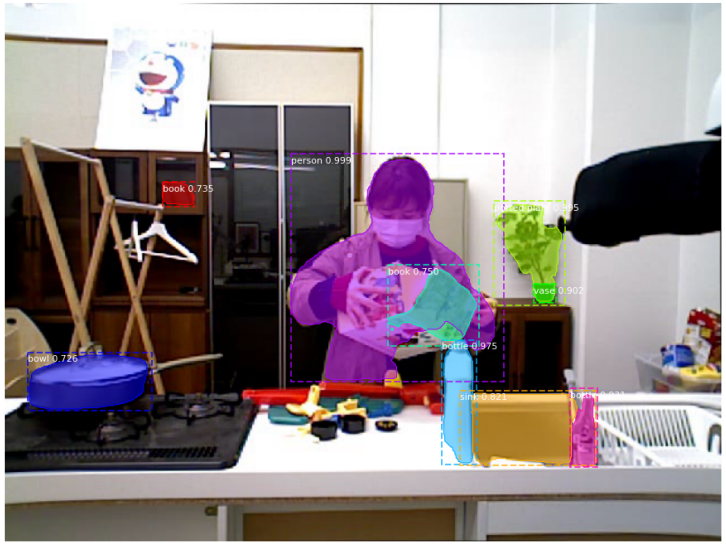
(a)



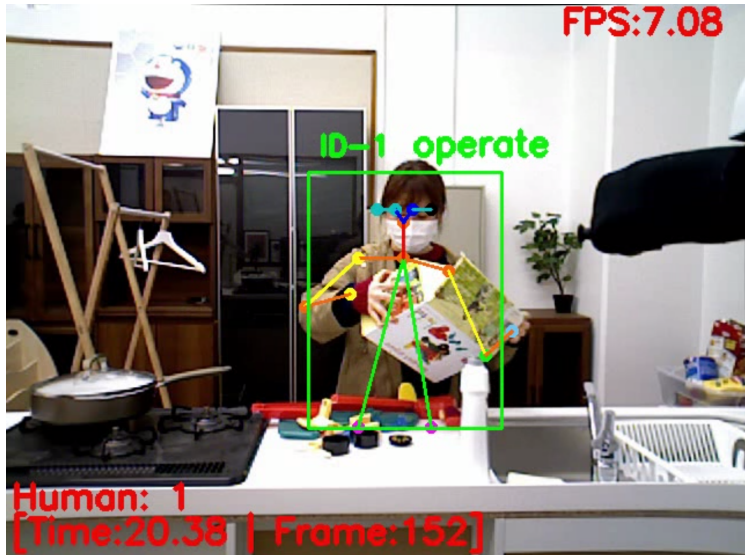
(b)

a) Segmentación de objetos en reconocimiento, b) Detección de esqueletos.

Resultados Parciales



(a)



(b)

a) Segmentación de objetos en reconocimiento, b) Detección de esqueletos.

Reporte de avances

Reporte de avances

Actividad	Estado	Tiempo Estimado
Colección de conjunto de datos	Completado	—-
Etiquetado de los datos	Completado	—-
Reentrenamiento de la red neuronal **(Reconocimiento de acciones)**	Completado	—-
Reentrenamiento de la red neuronal **(Detección de objetos)**	Completado	—-
Planteamiento de nuevo modelo inclusivo **(Detección Objetos + Reconocimiento de Acciones)**	Completado	—-
Entrenamiento y pruebas de nuevo modelo inclusivo **(Detección Objetos + Reconocimiento de Acciones)**	Completado	—-
Redacción de Tesis	En Proceso (70 %)	2 semanas

Código

- https://github.com/edgarVazquez43/action_recognition