Universidad Nacional Autónoma de México

Posgrado en Ciencias e Ingeniería de la Computación

Examen de Grado
Maestría en Ciencias e Ingeniería de la Computación

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

Edgar de Jesús Vázquez Silva

Tutor: Dr. Jesus Savage Carmona

7 de mayo de 2021

Contenido

Definición del problema

Motivación

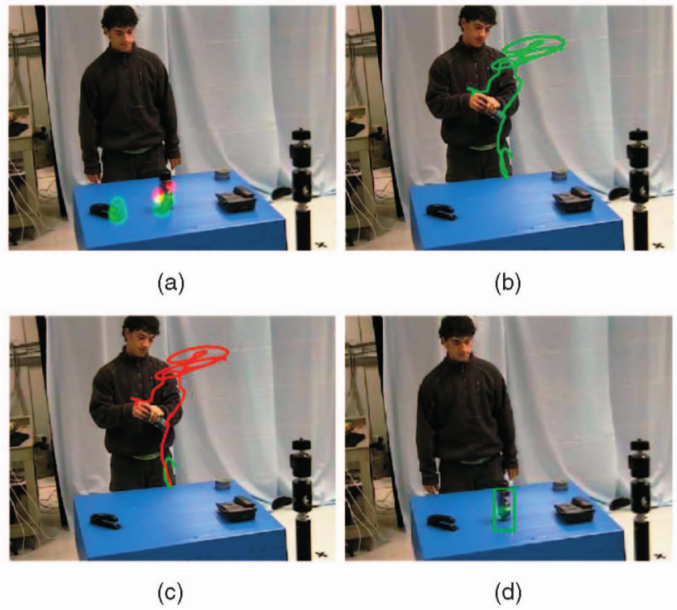


Figura 1. Interacción humano objeto. (a) Segmentación de objetos similares a una botella, (b) Trayectoria de manipulación de objetos, (c) Trayectoria de interacción con objeto, verde el agarre del objeto, rojo la manipulación de uso. (d) Categorización del objeto botella despúes de la interacción.

Definición del problema

- 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**: Si existe una relación entre la acción realizada por un humano y las características del objeto involucrado en tal acción, es posible inferir tal relación utilizando redes neuronales profundas.
- **Hipotesis 2**: Si conocemos la relación entre la acción realizada por un humano y las características del objeto involucrado en tal acción, se puede inferir propiedades sobre los objetos.

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.

Marco téorico

Redes Neuronales

Con base en la inspiración biológica del cerebro y los avances de la psicología de los años 50 surgió la hipótesis de poder simular los elementos básicos que procesaban información en el cerebro, resultado de esos estudios fue el **modelo de neurona artificial** [MP17].

Perceptrón

La definición formal del perceptrón de capa única se muestra en la ecuación 1, donde $H(x)$ es la función activación definida como una función escalón aplicada al resultado o:

$$H(x) = \begin{cases} 1 & si \quad \vec{w} \cdot \vec{x} + \omega \geq 0 \\ 0 & si \quad \vec{w} \cdot \vec{x} + \omega < 0 \end{cases} \tag{1}$$

En este caso se utiliza la función escalón como función de activación, sin embargo en actualidad se utilizan diferentes funciones de activación: sigmoide, relu, tanh, etc.

Perceptrón - Inspiración Biológica

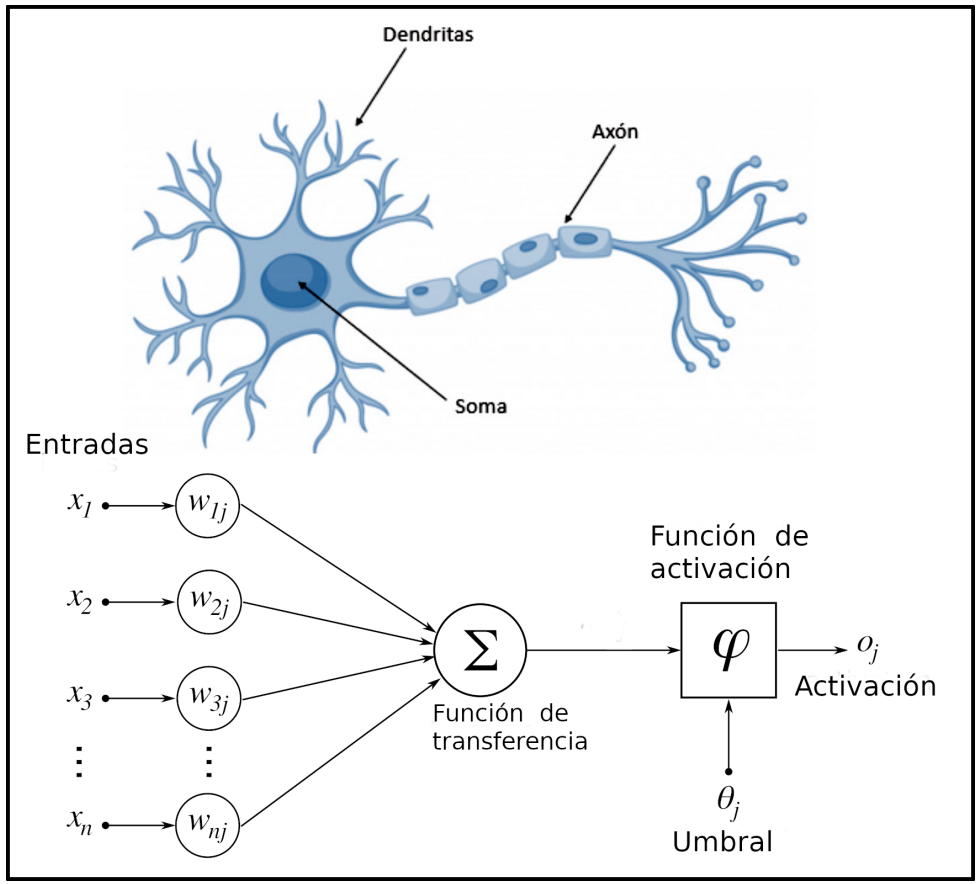


Figura 2. Modelo simplificado de de la inspiración biológica del modelo de neurona artificial.

Redes Neuronales Profundas

- Al abordar tareas cada vez complejas se aumentó el número de capaz y neuronas en el modelo neuronal.
- Algunas de las tareas promotoras de este incremento fue el reconocimiento de digitos escritos a mano [HS97].
- En 1998, Yann LeCun combinó por primera vez los modelos de convolución con algoritmos de retropropagación de errores.

Reconocimiento de Acciones

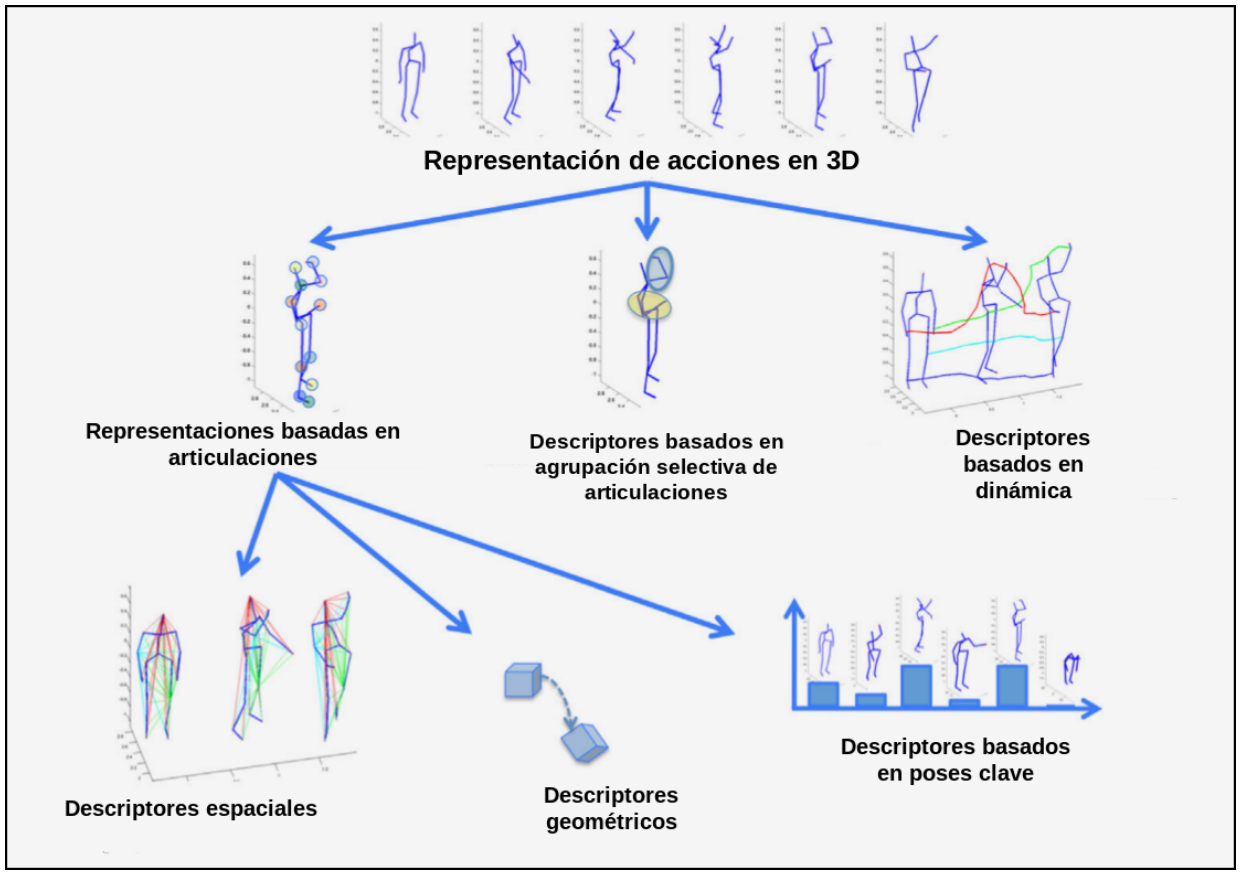


Figura 3. Esquema representativo de diferentes métodos de reconocimiento de acciones humanas basadas en esqueletos.

Detección y Reconomiento de objetos

- Dataset MNIST 1998, reconocimiento de cifras escritas a mano.
- Aparición de diferentes arquitecturas basadas en convoluciones RESNET, AlexNet, VGGNet, etc.

Metodología propuesta

Metodología propuesta

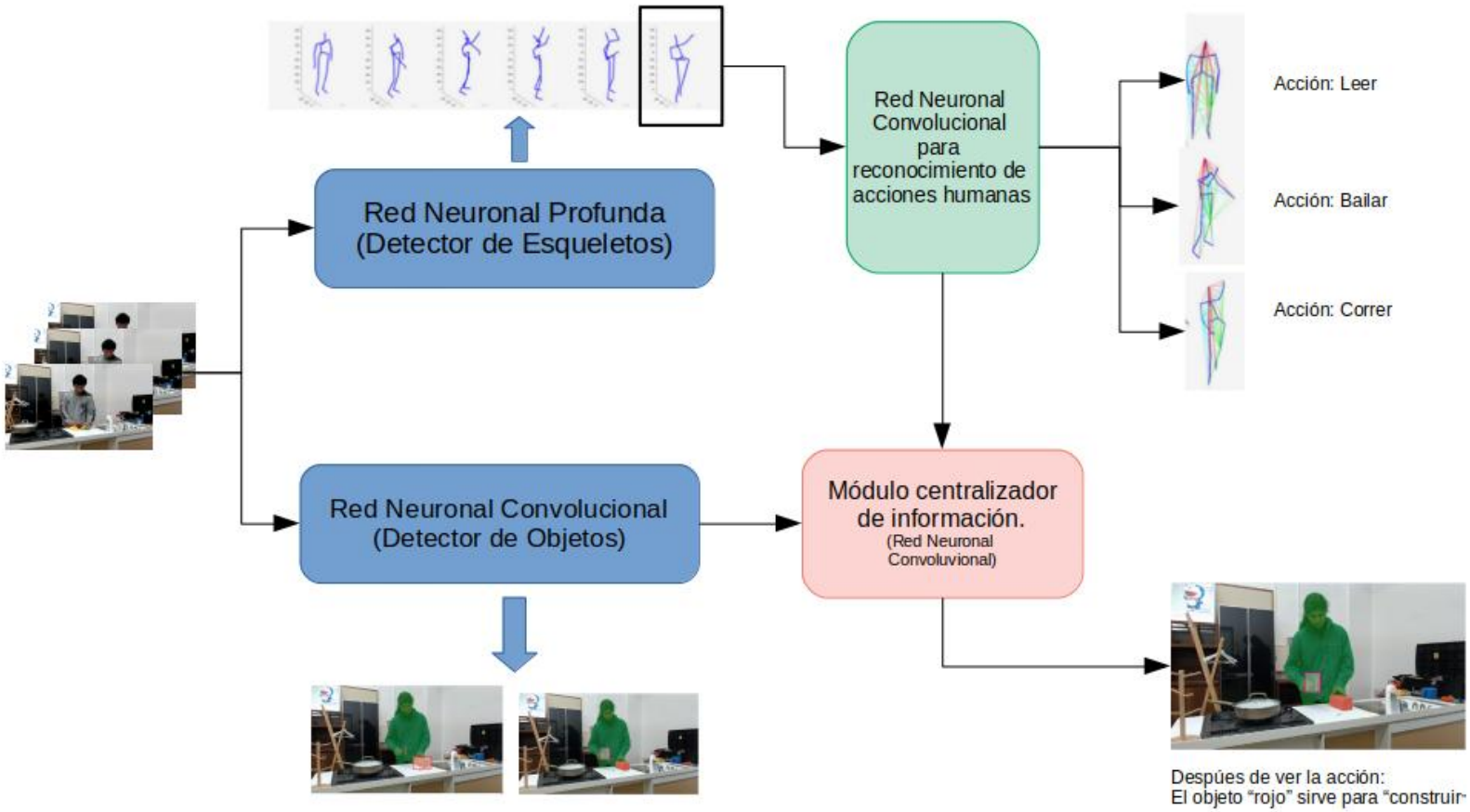


Figura 4. Descripción del sistema propuesto para la detección de interacción humano-robot.

Metodología propuesta

- Recolectar un conjunto de datos en escenas de entornos domésticos de prueba que sirva como base para entrenamiento y pruebas 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. (Imágenes RGB)
  - Sensor Xtion. (Imágenes RGB)
  - Camaras componentes del Robot HSR. (Imágenes RGB)
  - HD webcam C920 PRO. (Imágenes RGB)

Conjunto de datos (Dispositivos de Captura)

(a) Logitech webcam.

(b) Sensor kinect

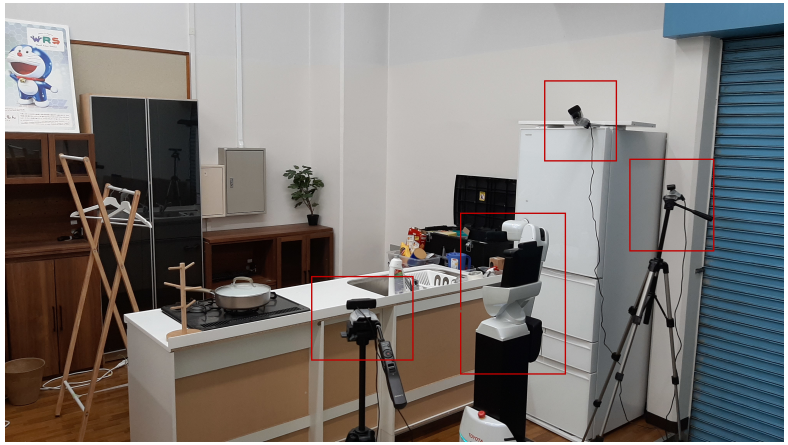
(c) Sensor Xtion Live Pro.

(d) Robot de servicio Doméstico HSR.

Figura 5. Dispositivos de captura del conjunto de datos

Conjunto de datos (Disposición de camaras)

(a)



(b)

Figura 6. 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.
- El proceso de etiquetado tardó aproximadamente 5 semanas.
- Para el proceso de reconocimiento de acciones se obtuvieron los esqueletos de cada persona que aparece en la escena, este proceso fue automatizado haciendo uso de la red neuronal **OpenPoses**.

Naturaleza de la información

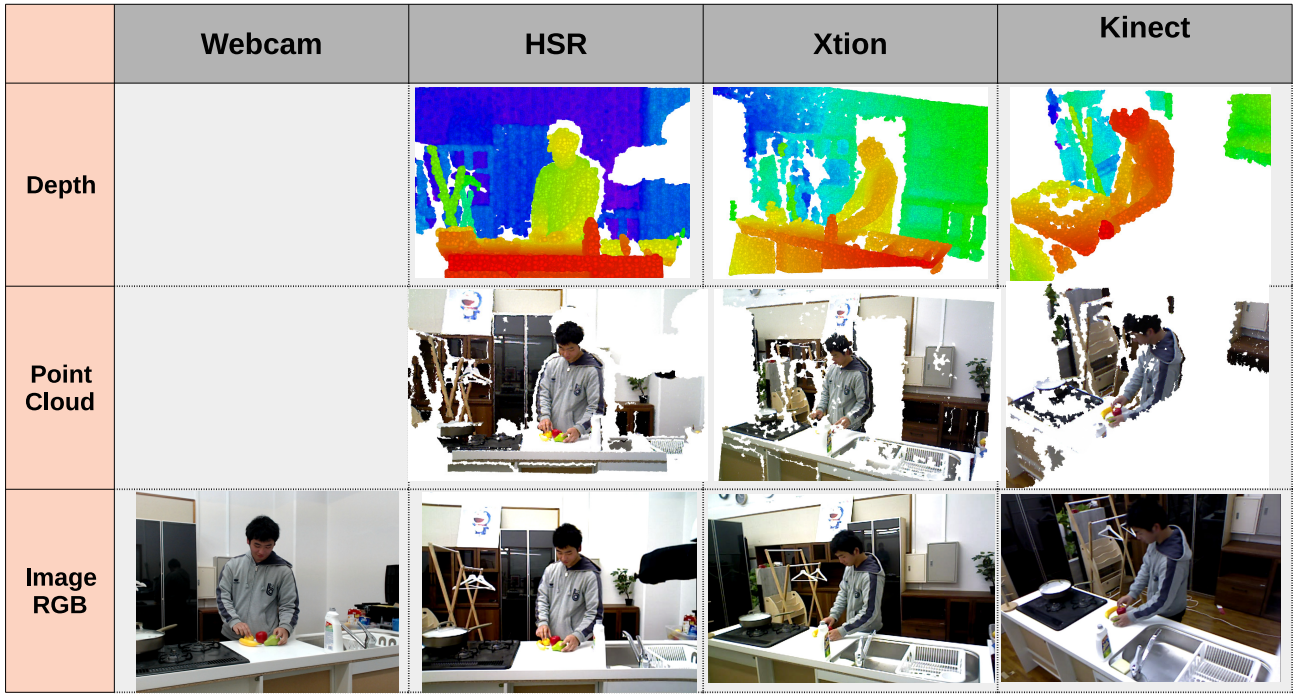


Figura 7. Ejemplo de información contenida en el Dataset. Muestra el tipo de información contenida de acuerdo al dispositivo de captura.

Conjunto de datos (Ejemplos)

(a) (b) (c)

Figura 8. 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)

Figura 9. Ejemplo de información contenida en el Dataset YCB.

Conjunto de datos (Etiquetado)

(a) (b) (c)

Figura 10. Etiquetado de objetos de interes en la escena. Para ello se hizo uso de la herramienta en línea *supervisely*.

Conjunto de datos (Etiquetado)

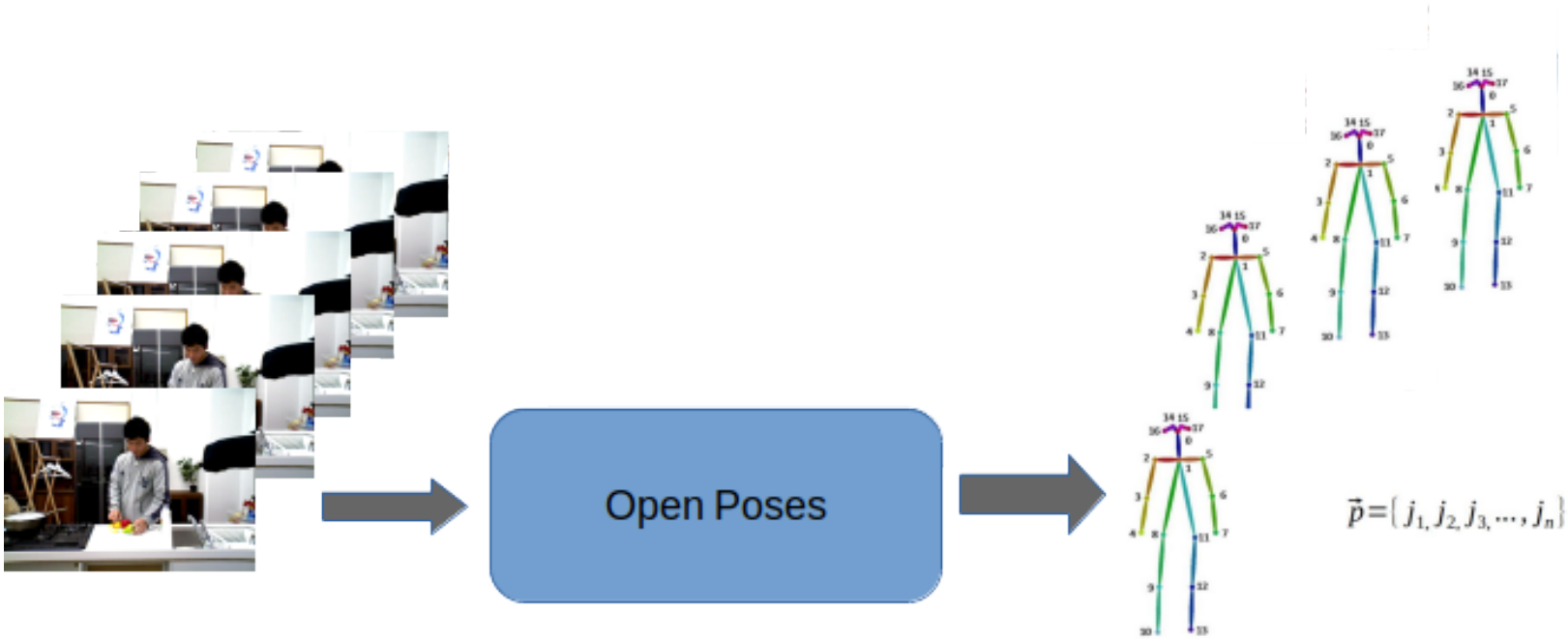


Figura 11. Generación de esqueletos a partir de imagenes, haciendo uso de la red **OpenPoses**.

Conjunto de datos (Etiquetado)

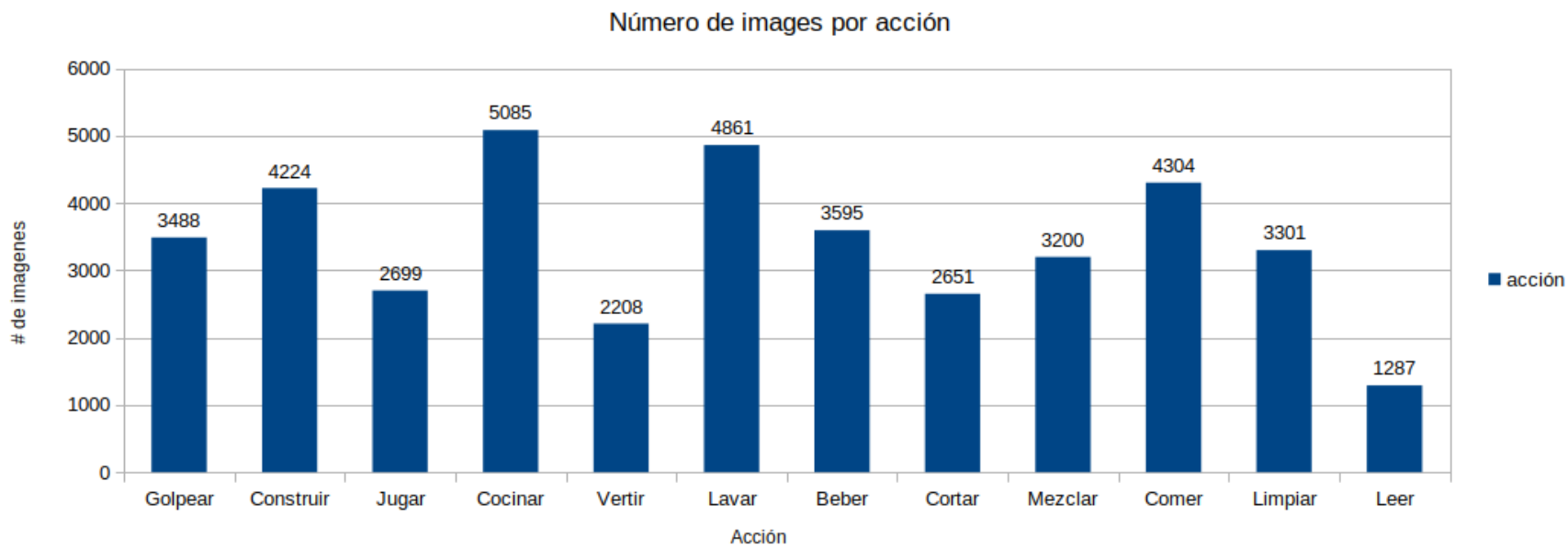


Figura 12. Total de datos etiquetados para cada una de las 12 acciones a detectar.

Sistema Propuesto

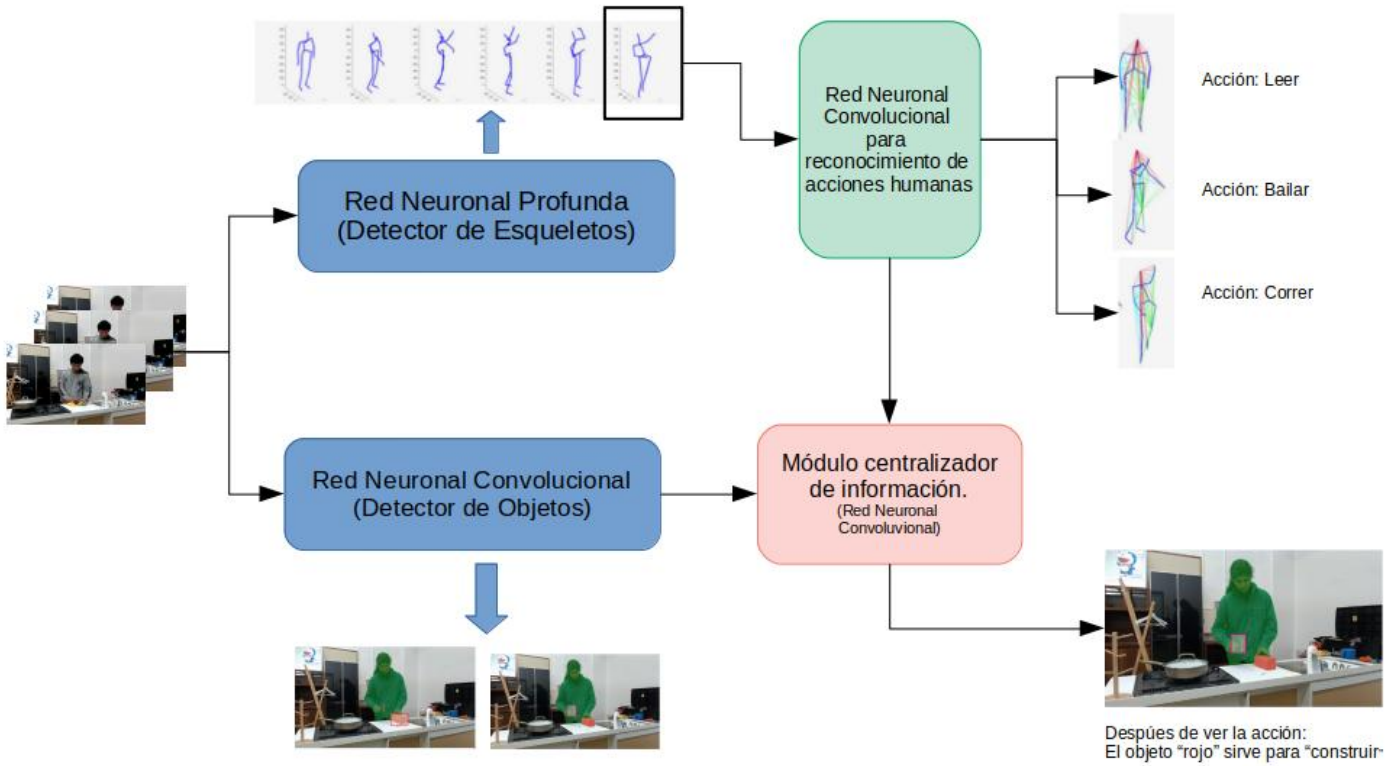


Figura 13. Descripción del sistema propuesto para la detección de interacción humano-robot.

Reconocimiento de acciones - Local

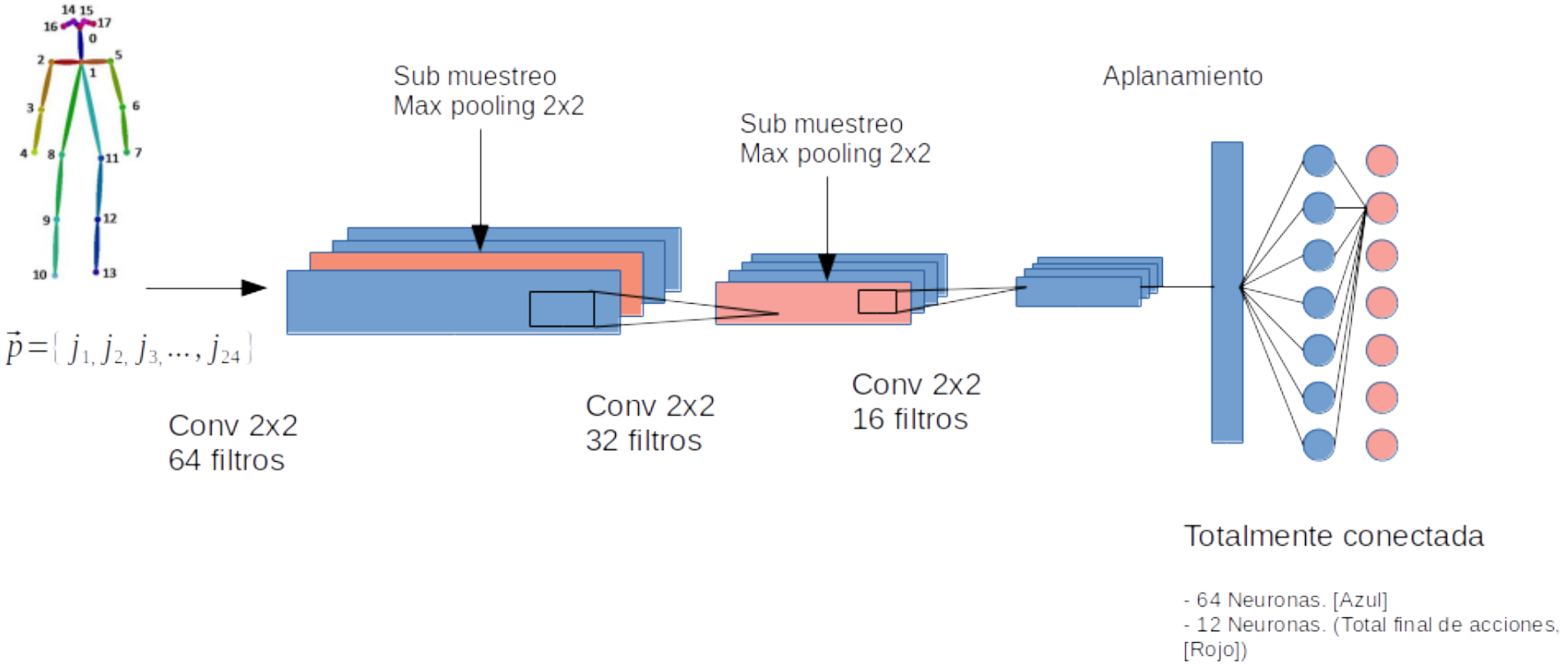


Figura 14. Modelo de reconocimiento de acciones usando unformación local.

Reconocimiento de acciones - Local

- Se utilizó OpenPose como extractor detector de esqueletos.
- Se abordó el problema desde un enfoque de clasificación de poses.
- 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 12 neuronas con una función de activación *softmax*.

Reconocimiento de acciones - Temporal

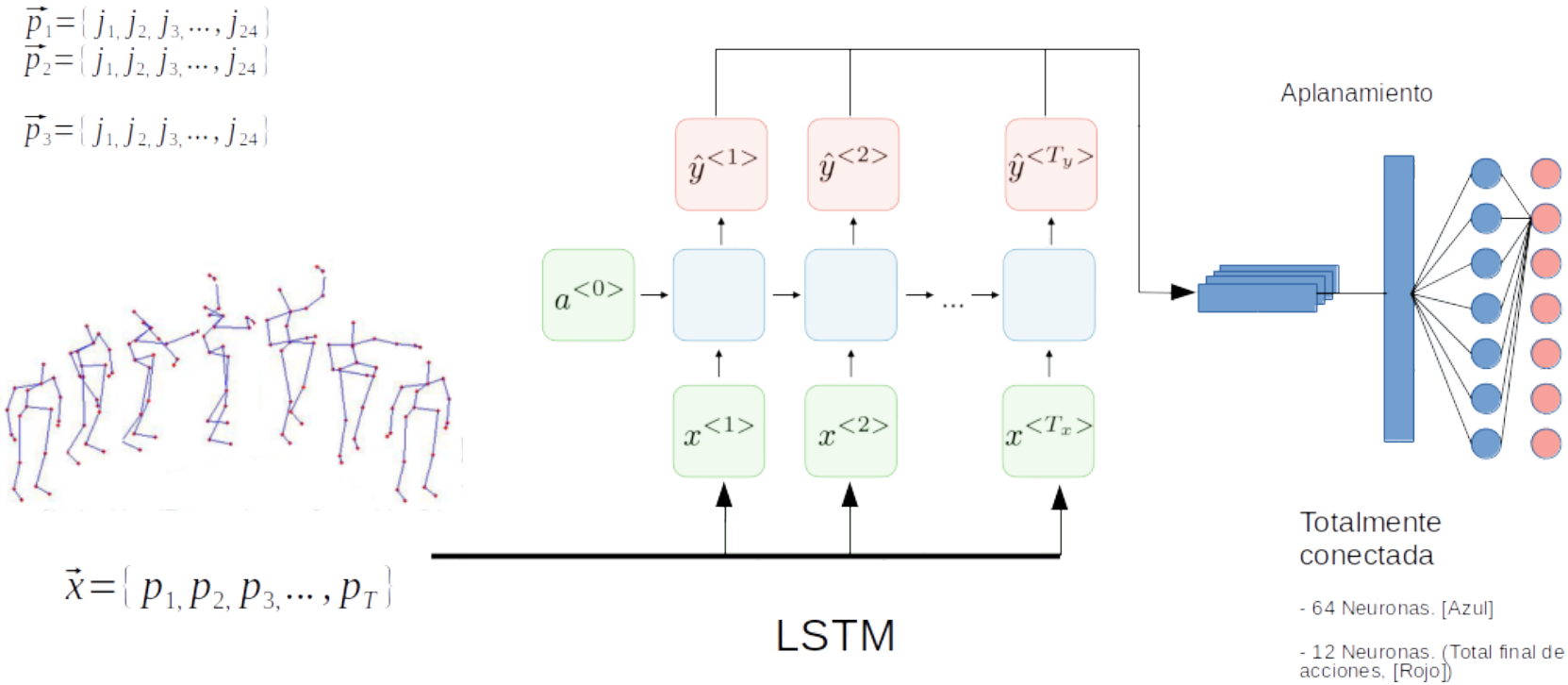


Figura 15. Modelo de reconocimiento de acciones usando unformación temporal

Reconocimiento de acciones - Temporal

- Se realizaron experimentos con un total de 120 poses consecutivas, 4 segundos de vídeo muestreados a 30 fps.
- Se hizo uso de una sola capa de neuronas LSTM con 50 elementos.
- Dentro de cada celda un vector de estados es compartido con la siguiente unidad, este vector de estados es de tamaño 255.
- En la etapa final se toman los estados de cada elemento LSTM y se pasan por un clasificador multi-clase.

Detección de Objetos

- Se utilizó la red neuronal Mask-RCNN, una red de segmentación semántica.

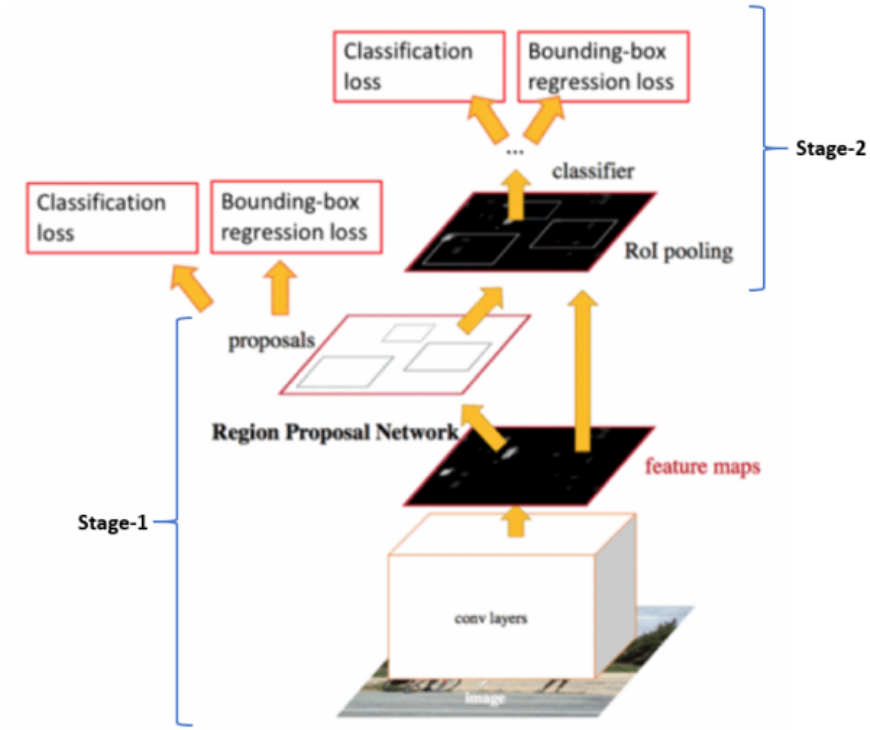
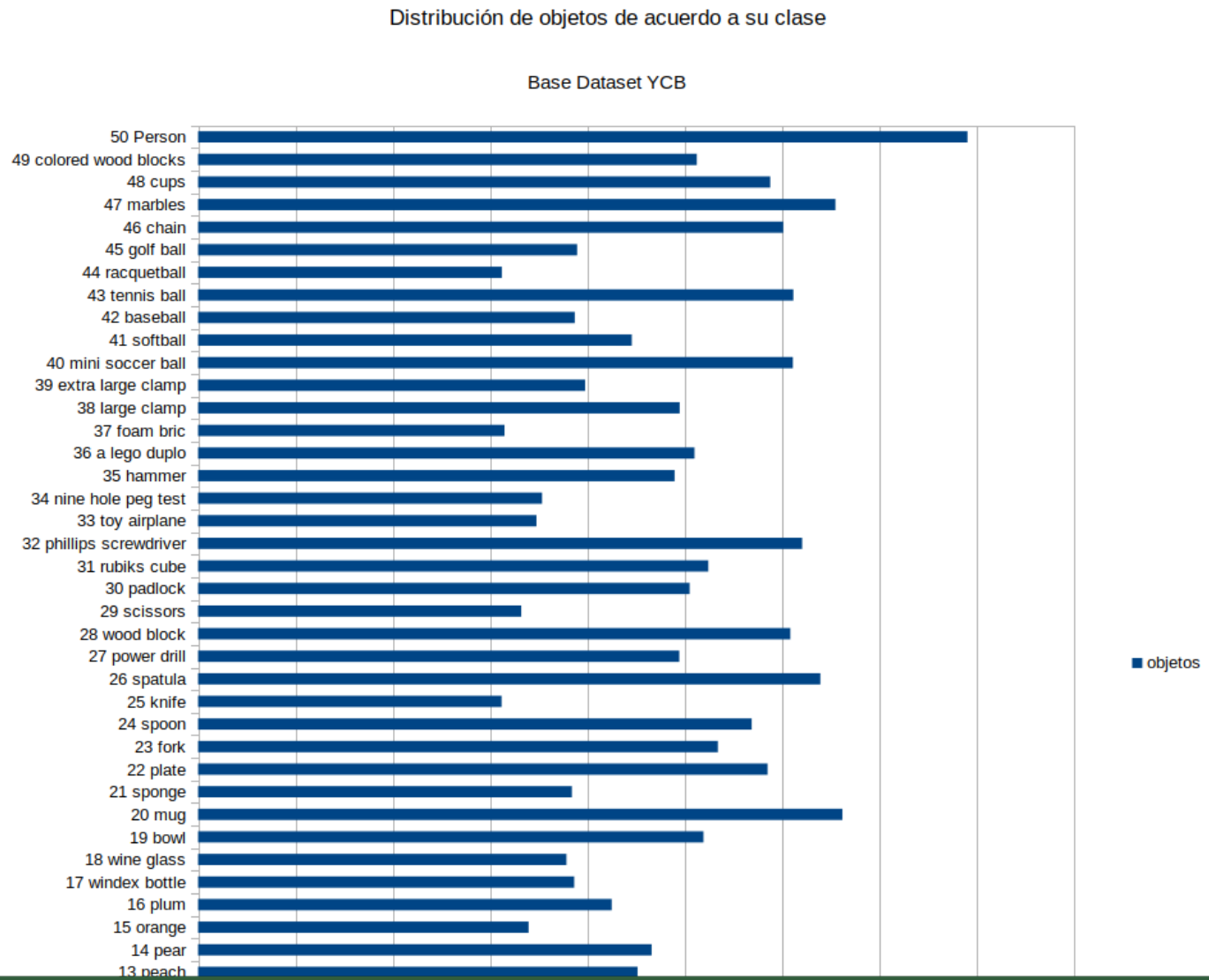


Figura 16. Esquema simplificado de la red Mask-RCNN [HGDG17].

Detección de Objetos

- Se utilizó la red pre-entrenada en el dataset COCO (Common Objects in COntext).
- Se realizó *transferencia de conocimiento* con el dataset limitado para esta tarea.
- Se utilizaron 50 objetos diferentes pertenecientes al dataset YCB (Yale-Carnegie Mellon-Berkley).

Detección de Objetos



Propiedades de uso de los objetos

Acción	Inferencia en uso del objeto en cuestión
Golpear	Herramienta
Construir	Herramienta
Jugar	Juguete
Cocinar	Utensilio de cocina
Vertir	Contenedor
Lavar	Articulo de limpieza
Beber	Contenedor / Bebida
Cortar	Utensilio / Cubiertos
Mezclar	Utensilio de Cocina
Comer	Comida
Limpiar	Articulo de limpieza
Leer	Pasa tiempo

Tabla 1. En esta tabla se muestra la relación existente entre la acciones humanas y las propiedades de uso de los objetos en cuestión.

Integración

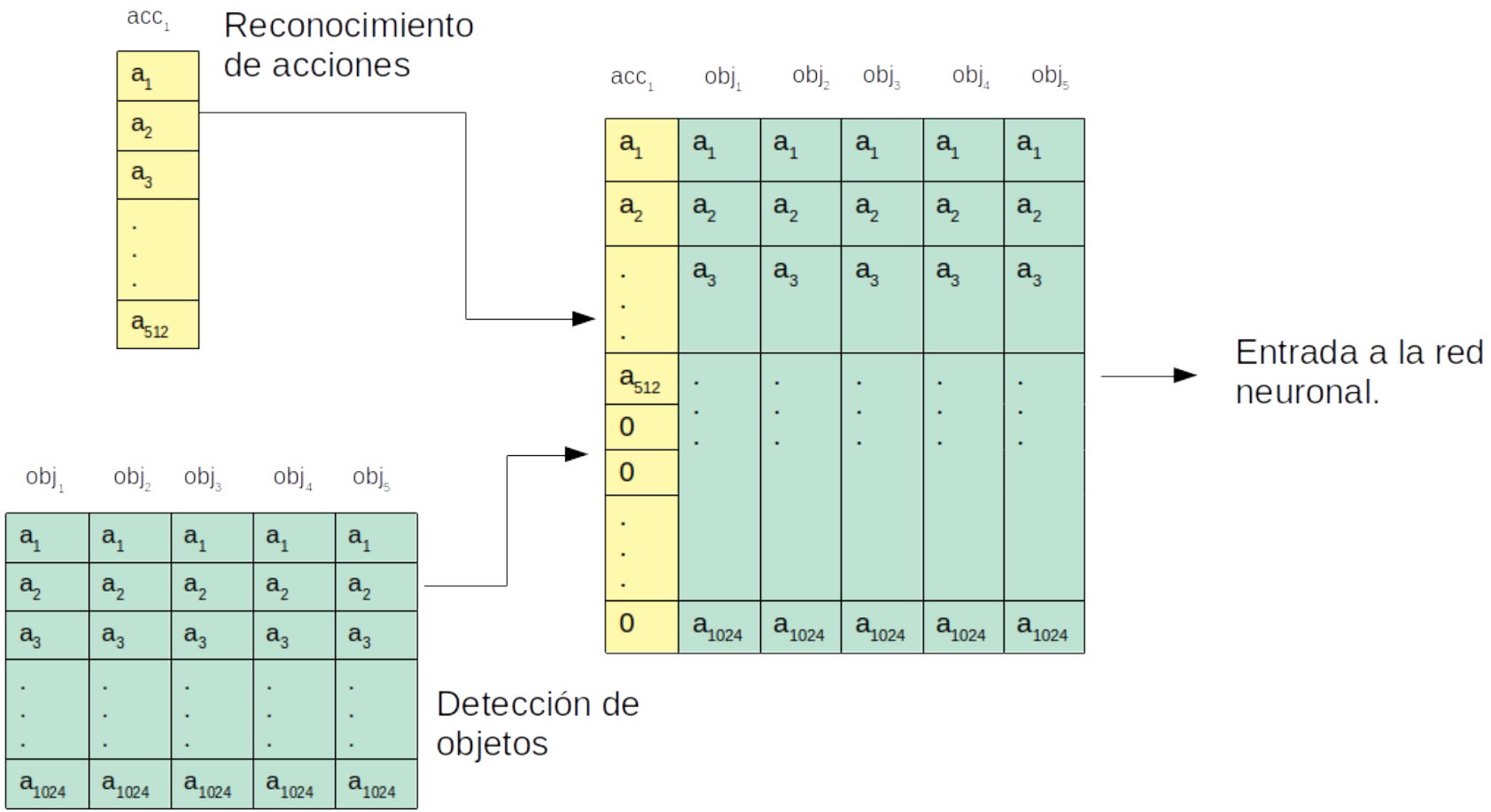


Figura 18. Modulo predictor de propiedades.

Integración

- Modulo encargado de unificar la información de las acciones con la información de los objetos.
- Posee una arquitectura de red neuronal convolucional que toma como entrada una matriz de 6x1024.
- La arquitectura consta de 2 capas convolucionales con una capa de sub-muestreo correspondiente. Y una última capa de clasificación.

Integración

- El sistema en su conjunto se habilitó en dos computadoras conectadas en una red local, haciendo uso de ROS (Robot Operative System).

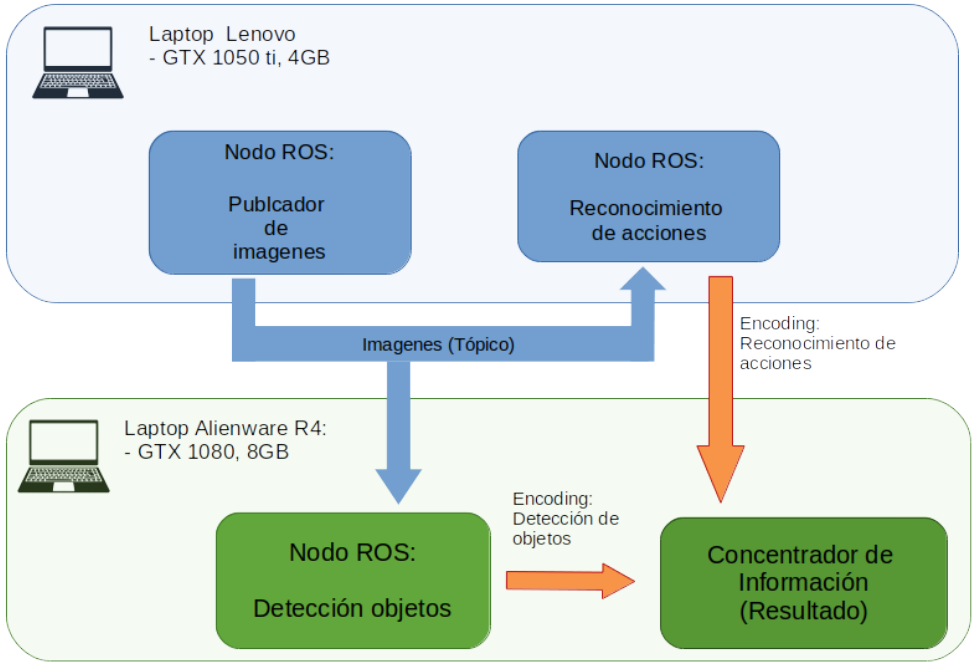


Figura 19. Esquema representativo de los módulos componentes del sistema, se muestra el sistema distribuido en diferentes computadoras.

Resultados

Reconocimiento de acciones - Local

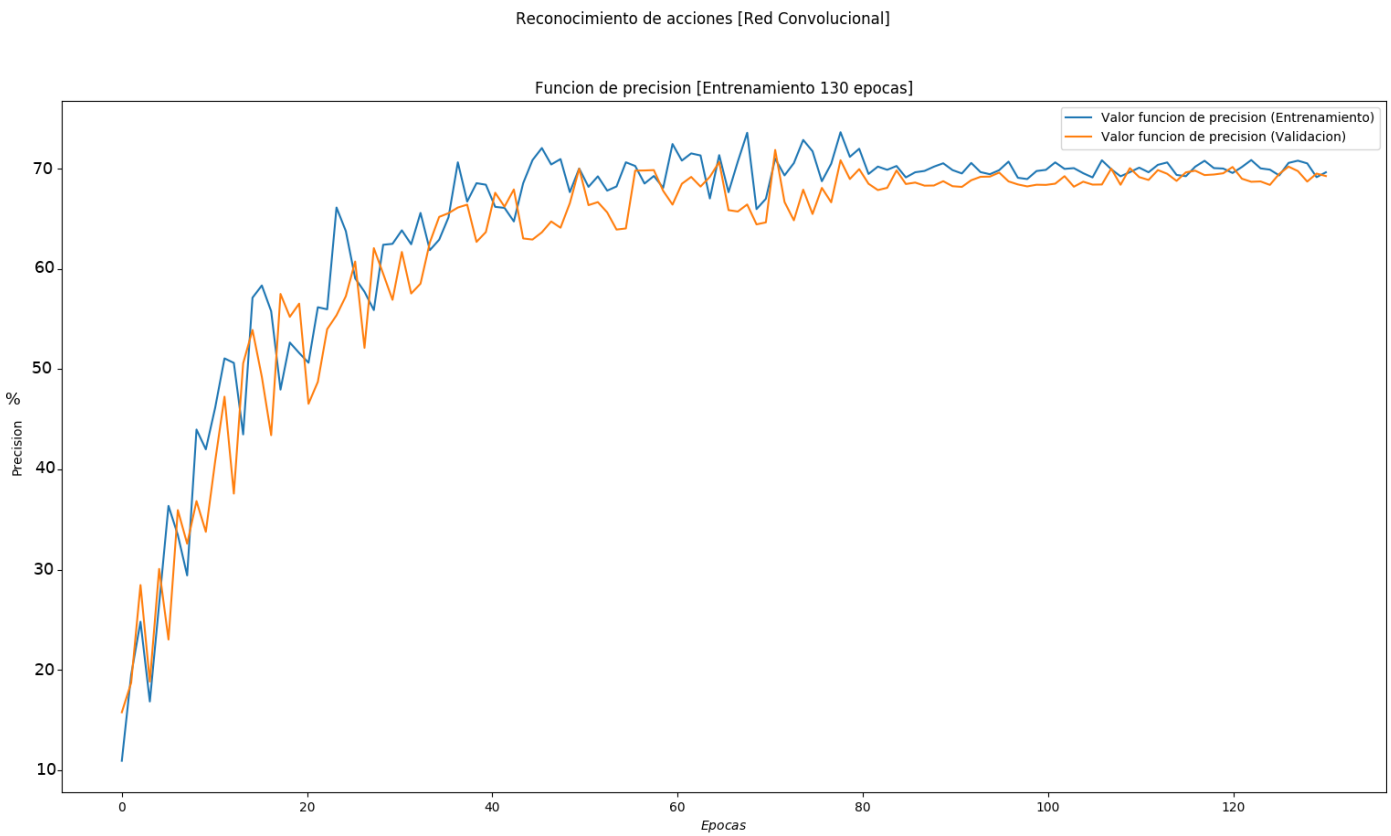


Figura 20. Gráfica de exactitud del modelo de reconocimiento de acciones con información local para 130 epocas.

Reconocimiento de acciones - Local

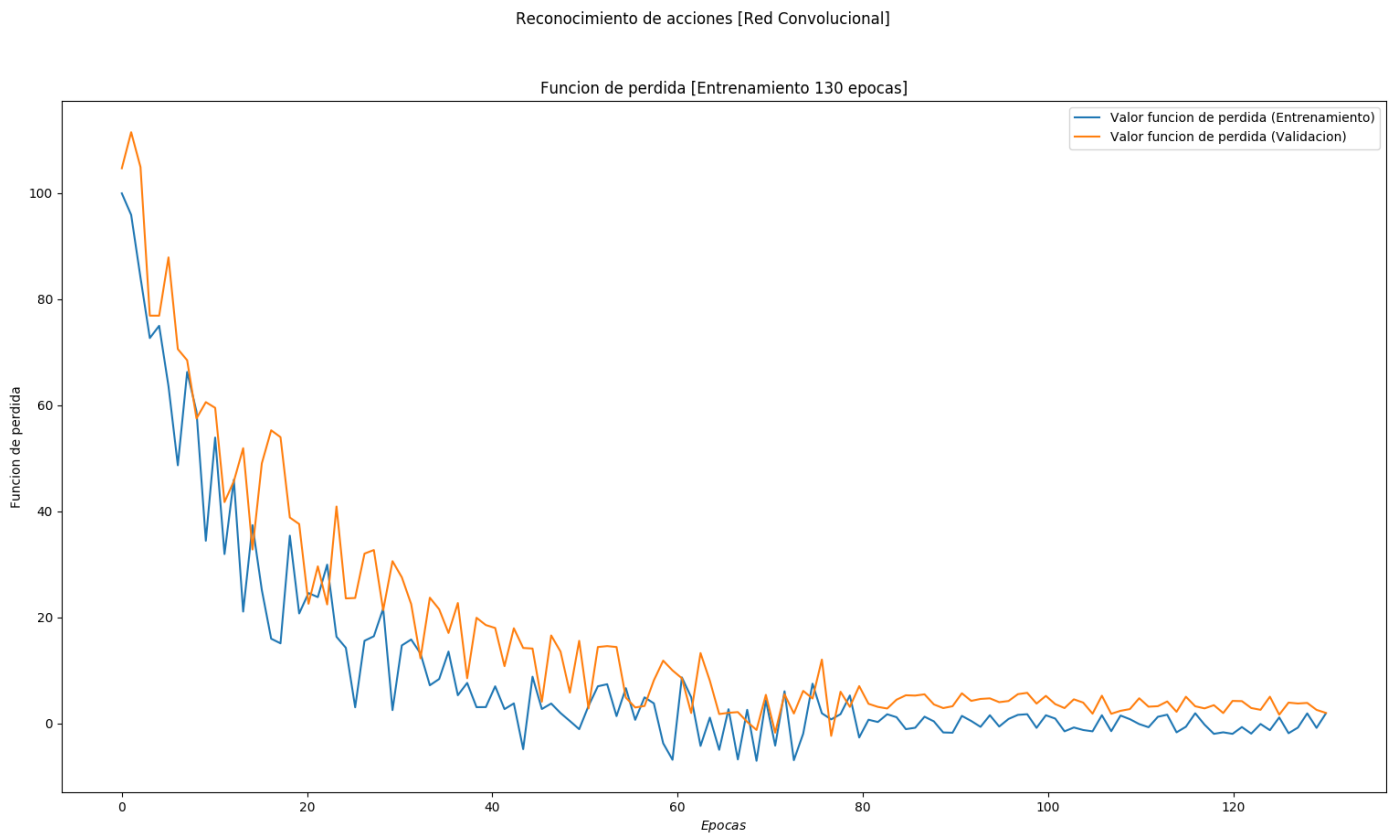


Figura 21. Gráfica de la función de perdida para el modelo de reconocimiento de acciones con información local para 130 epocas.

Reconocimiento de acciones - Recurrente

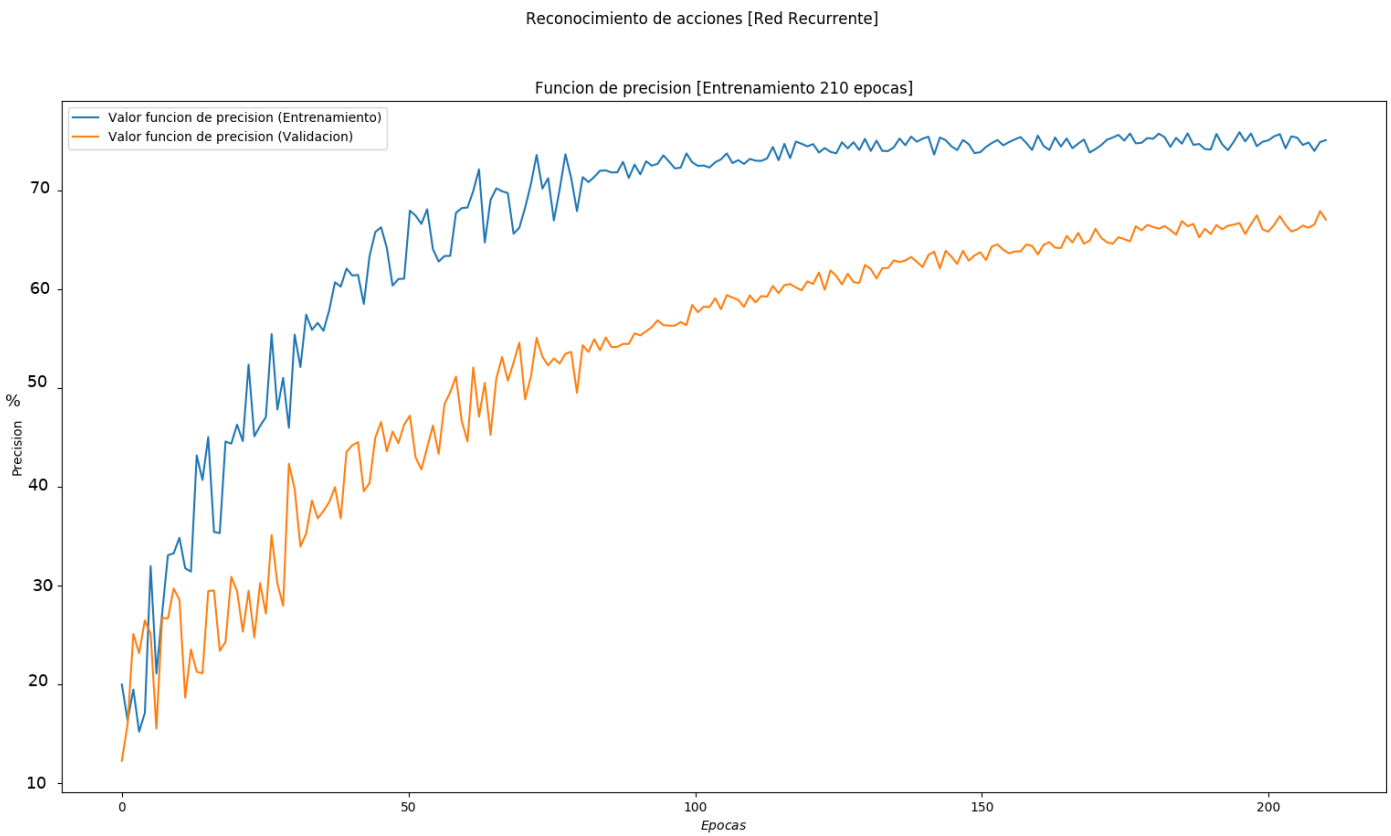


Figura 22. Gráfica de exactitud del modelo de reconocimiento de acciones con información temporal para 210 epocas.

Reconocimiento de acciones - Recurrente

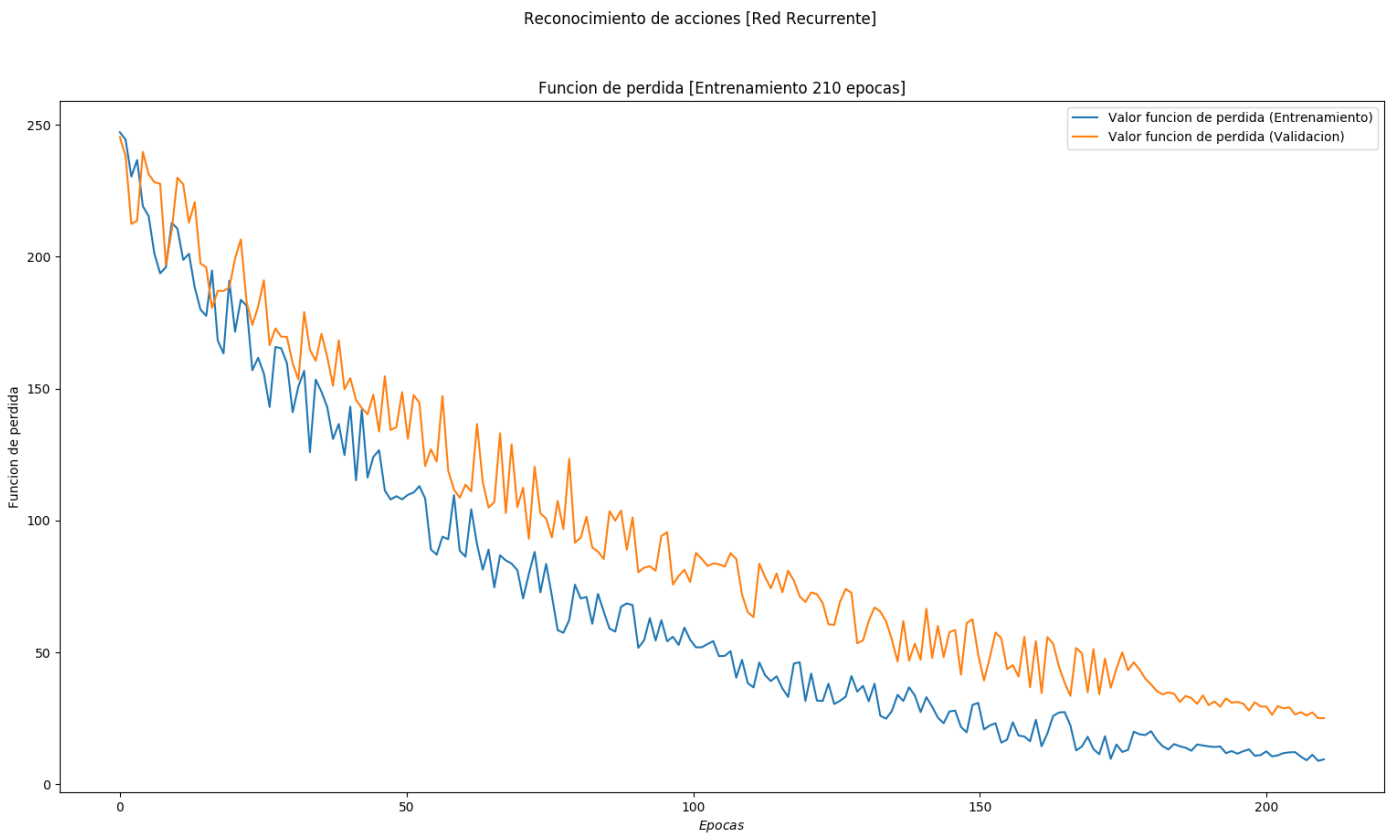


Figura 23. Gráfica de la funcion de perdida del modelo de reconocimiento de acciones con información temporal para 210 epocas.

Reconocimiento de acciones Conv2D

- Número de parametros: 12,004,485
- Tiempo de entrenamiento: 14 horas 20 minutos (Nvidia 1050GTX)
- Porcentaje final de exactitud en validación: 0.68
- Tiempo en inferencia: 0.2314 [s]

Reconocimiento de acciones RNN

- Número de parametros: 34,663,525
- Tiempo de entrenamiento: 18 horas 34 minutos (Nvidia 1050GTX)
- Porcentaje final de exactitud en validación: 0.708
- Tiempo en inferencia: 0.8853 [s]

Matriz confusión (Reconocimiento de Acciones)

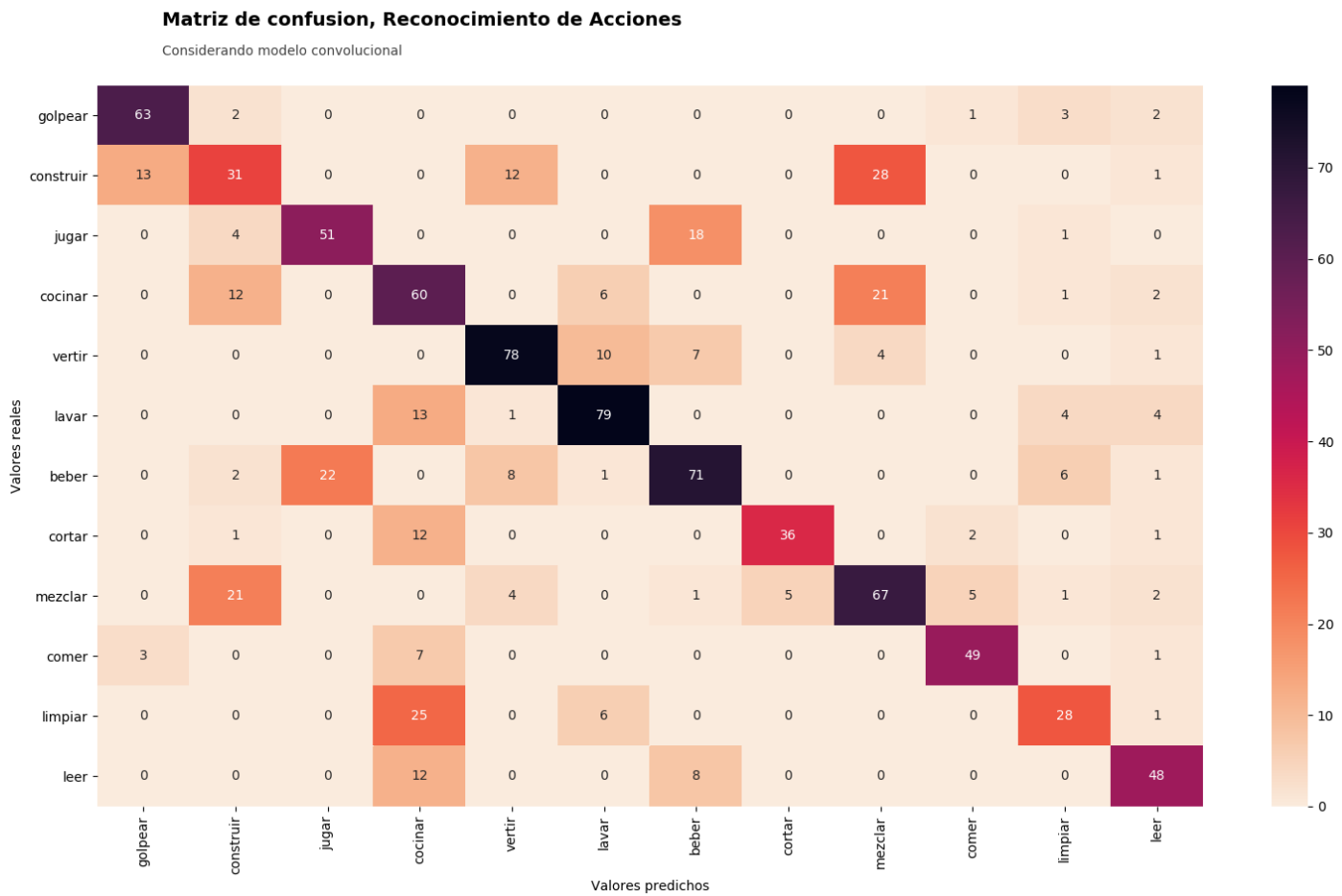


Figura 24. Matriz de confusión para el modelo de reconocimiento de acciones

Matriz confusión (Reconocimiento de Acciones)

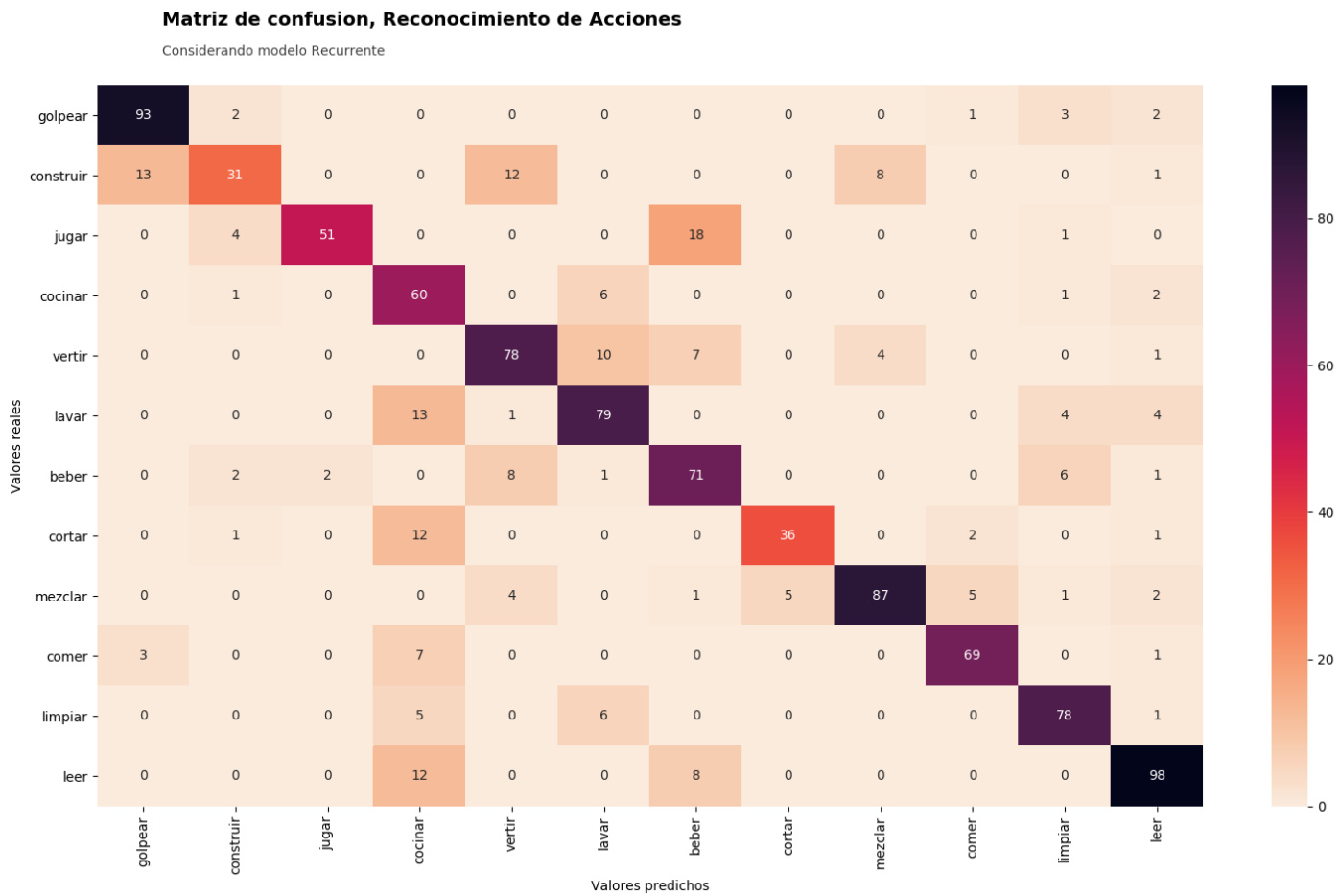


Figura 25. Matriz de confusión para el modelo de reconocimieno de acciones

Detección de Objetos

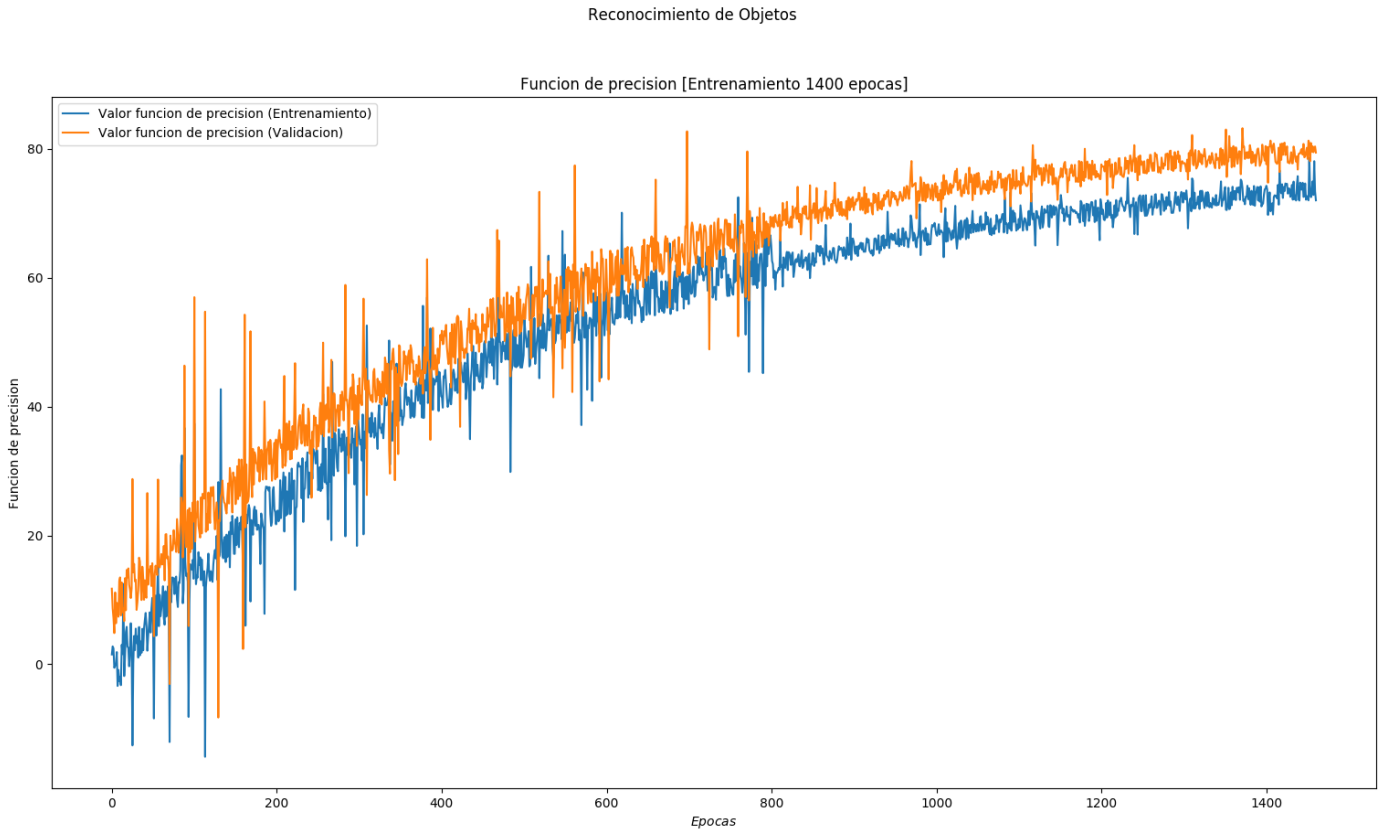


Figura 26. Gráfica de exactitud del modelo de reconocimiento de objetos.

Detección de Objetos

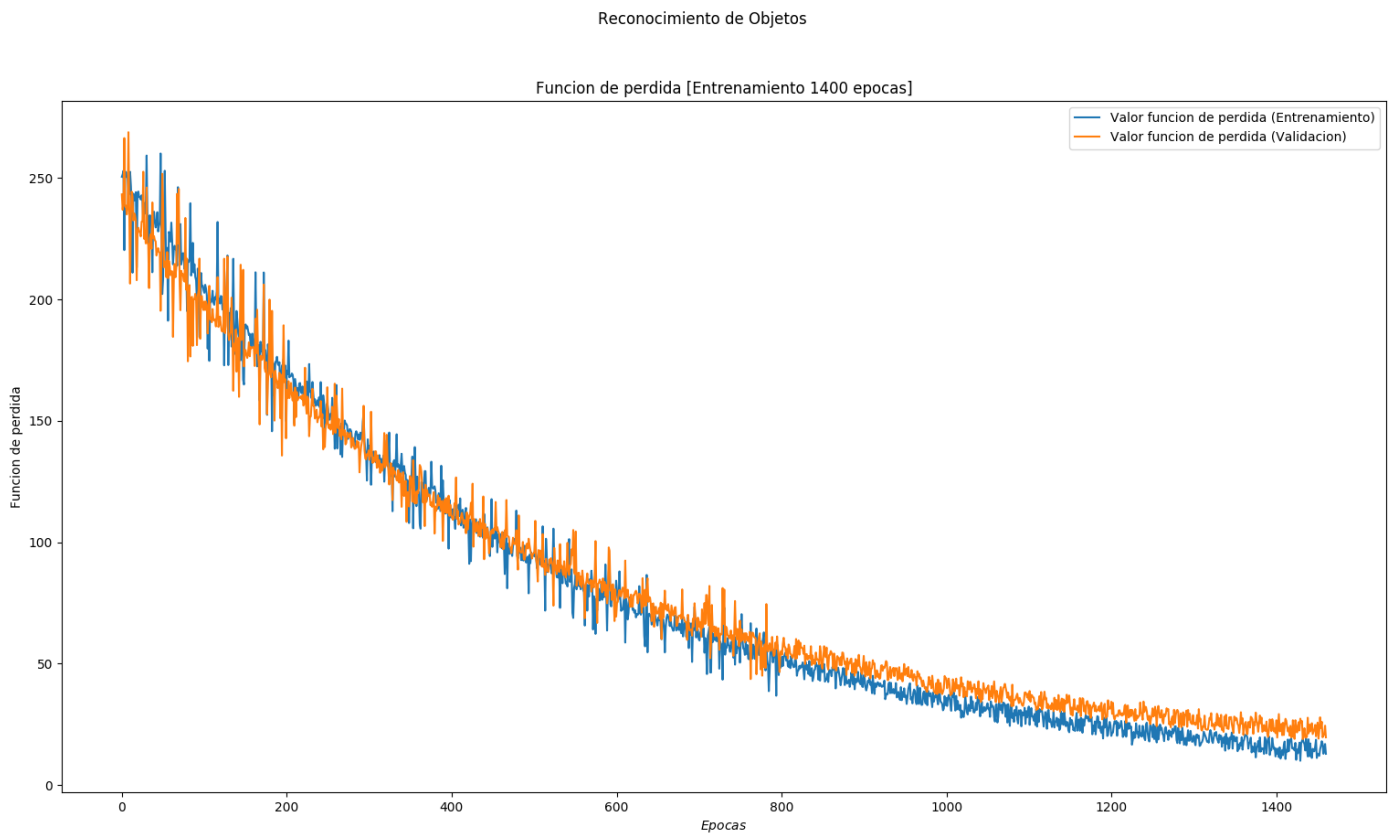
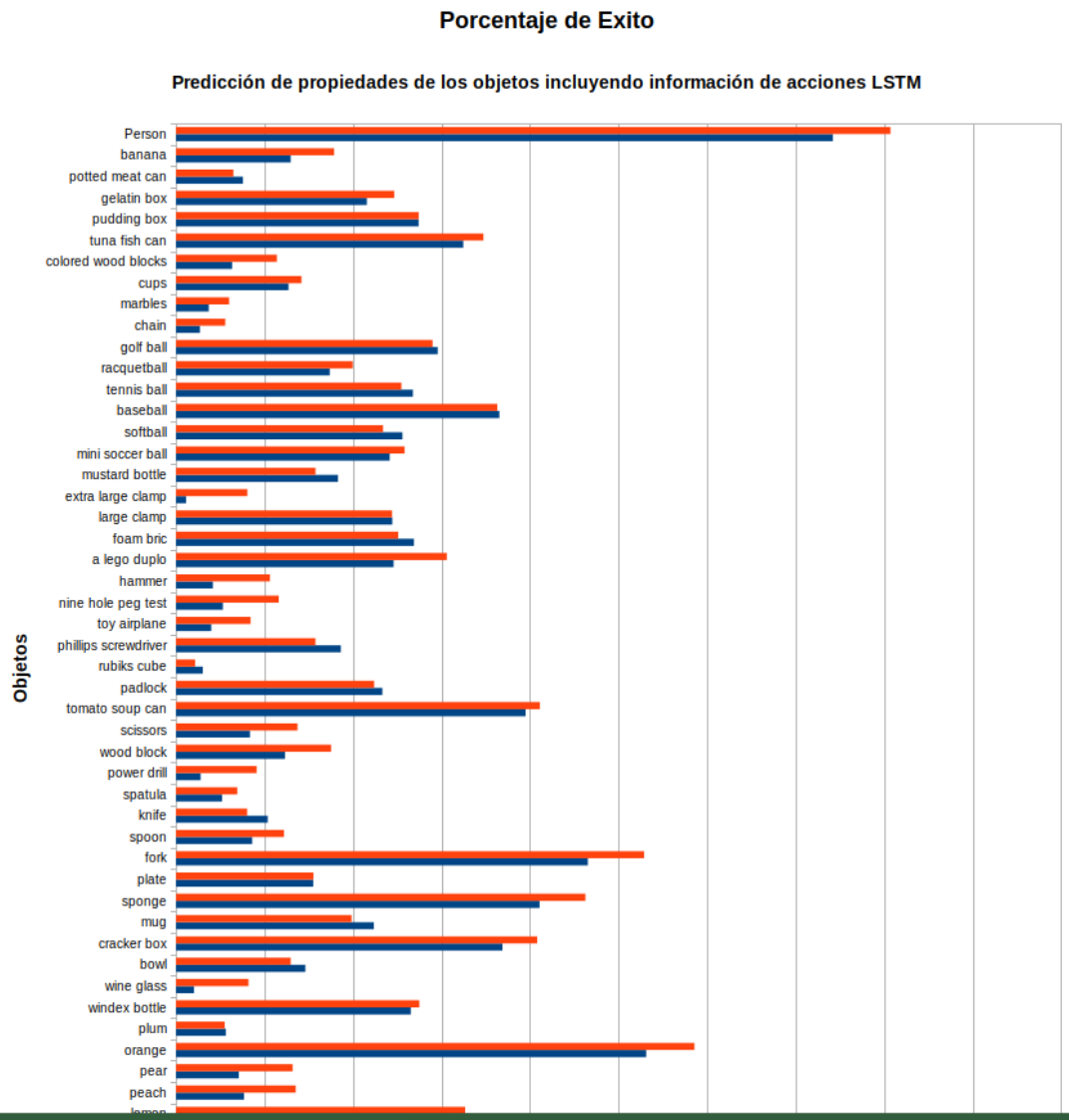


Figura 27. Gráfica de la funcion de perdida del modelo de reconocimiento de objetos.

Detección de Objetos

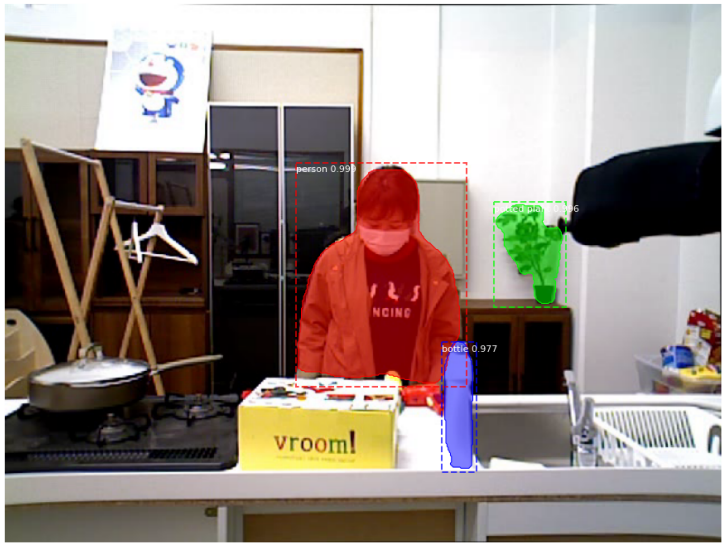


Prueba T para muestras pareadas

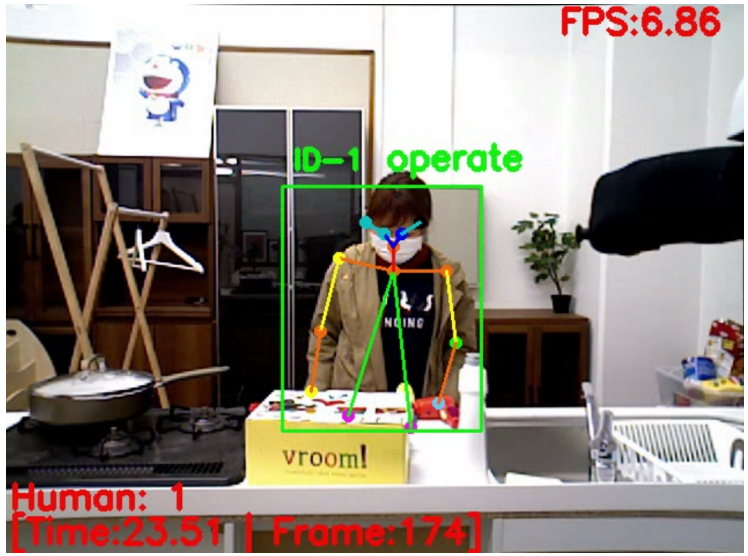
	Sin información de acciones	Con información de acciones
Media de exactitud	19.88580	22.19844
Desviación estándar	15.16316	15.43039
Valor - t	5.1475	
Valor - p	$4{,}656 \times 10^{-6}$	

Tabla 2. Tabla con los resultados de la prueba t-student para muestras pareadas aplicada a éxito de inferencia de propiedades sobre los objetos.

Resultados Parciales



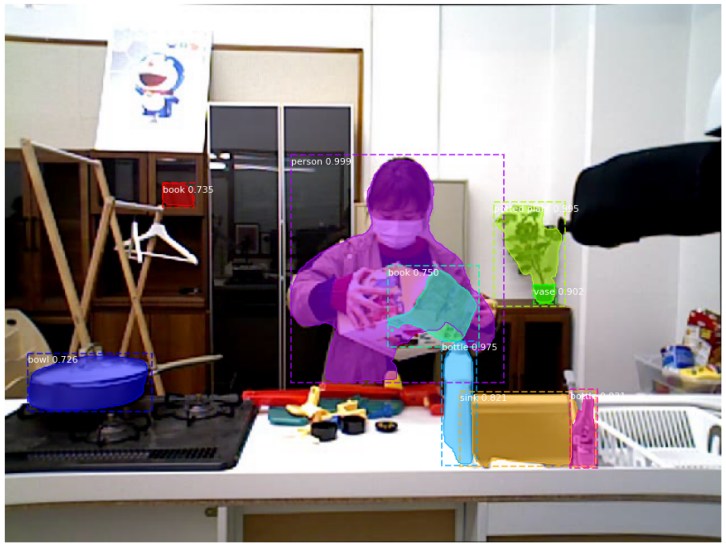
(a)



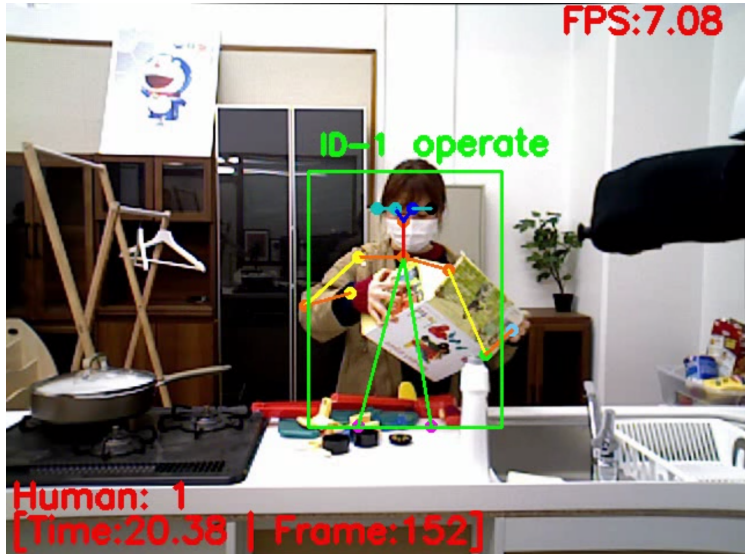
(b)

Figura 29. a) Segmentación de objetos en reconocimiento, b) Detección de esqueletos.

Resultados Parciales



(a)



(b)

Figura 30. a) Segmentación de objetos en reconocimiento, b) Detección de esqueletos.

Discusión

- Existe una gran complicación al etiquetar una gran cantidad de datos en el caso del aprendizaje supervisado de redes neuronales profundas.
- Existe una limitación al ampliar el número de acciones a reconocer debido a que el cuerpo humano solo es **parcialmente visible**.
- El dataset desbalanceado dificulta la comparación de exactitud entre diferentes clases de acciones.
- La naturaleza de las acciones humanas son de duración variable, lo cual no está representado fielmente en esta metodología.

Discusión

- El modelo con información temporal requiere un paso extra que es formar el buffer de entrada lo cual retrasa la respuesta de todo el sistema.
- La matriz de confusión para el modelo de reconocimiento de acciones humanas con información temporal presenta menor dispersión respecto a la línea diagonal con respecto del mismo caso para el modelo con información local.
- El reconocimiento de objetos, se puede mejorar utilizando un modelo pre-entrenado especialmente para el dataset YCB.
- Una posible solución puede ser el aumento de datos para lograr una generalización.

Conclusiones

Conclusiones Parciales (Reconocimiento de acciones)

- Ambos modelos neuronales convergen hacia el conjunto de datos de manera esperada. Figuras 21, 22, 23, 24.
- Se observó un mejor porcentaje de presición con el modelo de información temporal para el reconocimiento de acciones humanas. Lamina 42, 43.
- El modelo de reconocimiento de acciones humanas con información local es más rapido en inferencia para montar un sistema en línea. Lamina 43 .

Conclusiones Parciales (Reconocimiento de acciones)

- Los entrenamientos se realizaron con un dataset acotado, debido a esto los resultados en ambientes dinámicos reales o estandarizados de pruebas (Robocup), pueden no generalizar lo suficientemente bien.

Conclusiones Parciales (Reconocimiento de objetos)

- Los objetos que presentan un menor porcentaje de éxito en la detección y el reconocimiento son de un tamaño menor con respecto al resto de los objetos, esto es debido a que las cámaras no son de alta resolución.
- El objeto que mejor identifica el modelo entre el conjunto de los 50 objetos presentes es precisamente la figura humana.
- El modelo y entrenamiento tiene un desempeño pobre en comparación con otros modelos de detección y reconocimiento de objetos [RDGF16].

Conclusiones Parciales (Unión de Sistemas)

- Existe una mejora significativa al inferir propiedades de los objetos cuando se añade información de las acciones humanas al proceso. Tabla 2.

Se establece que el valor p resultado de la prueba t debe ser menor o igual que ($p \leq 0{,}01$) para poder concluir que existe significancia estadística entre ambas mediciones. En este caso en particular se tiene un valor $p = 4{,}656 \times 10^{-6}$.

Trabajo Futuro

Trabajo Futuro

- Unificar los diferentes modelos neuronales.
- Implementar el modelo en un Robot de Servicio Doméstico y obervar el comportamiento con información fuera del dataset de entreamiento.
- Hacer uso de la información de profundidad que se logró capturar en el dataset pero que no se usó en el presente trabajo.
- Abordar el problema con un tipo de red neuronal basado en atención, ya que estos modelos presentan un mejor desempeño en el reconocimiento de acciones según el estado del arte hasta el año 2019 [GCDZ19].

Bibliografía

Bibliografía I

- Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman, *Video action transformer network*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 244–253.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, *Mask r-cnn*, Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
- Sepp Hochreiter and Jürgen Schmidhuber, *Long short-term memory*, Neural computation **9** (1997), no. 8, 1735–1780.
- Marvin Minsky and Seymour A Papert, *Perceptrons: An introduction to computational geometry*, MIT press, 2017.

Bibliografía II

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, *You only look once: Unified, real-time object detection*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.