Visión Computacional, Aplicación y Estado de Arte

Medina Lopez ,Jahir. Pacheco Miranda ,Miguel.
jahir.medina@unitru.edu.pe — jpacheco@unitru.edu.pe
Facultad de Ciencias Fisica y Matematicas, Informatica

***Resumen*—El presente articulo recopila las aplicaciones y avances del área de la informática y electrónica denominada Visión Computacional, considerando como marco referente el estado de arte del mismo.**

Si bien en el pasado [mit,] la *V.C.*(siglas de visión computacional) se encontraba alejado del consumidor, es decir, las aplicaciones relacionadas al área tenían un factor académico u exploratorio; el distanciamiento era tal que la universidades debían tener departamentos especializados, haciendo prácticamente imposible una aproximación amateur a este campo de estudio.

En la actualidad, con el avance de la electrónica, el perfeccionamiento de métodos matematismos mas complejos y el mas común uso de frameworks y librerías especializadas [Keim et al., 2008] (OpenCV, CUDA, TensorFlow); los progresos en investigación rivaliza en velcidad con su capacidad para generar productos finales [maj,] (Google Glass, Microsft Hololens, Tesla Self Driving Cars).

Sin embargo, a mayor uso y mejor desempeño de las aplicaciones del *PDI* aparecen nuevos problemas técnicos.

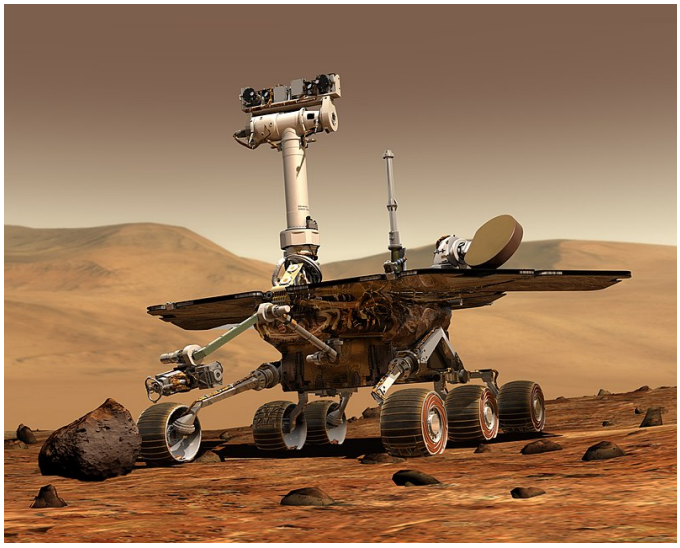

Figura 1. Curiosity Rover

***Index Terms*—Visión Computacional, Aprendizaje de Maquina, Técnicas no Supevisionadas, Realidad Aumentada**

I. Introducción

La visión computacional es el área que mas rápidamente a crecido en los últimos aos, en términos de aplicación e investigación, si bien es cierto parte de su desarrollo esta relacionado con el avance de las rede neuronales artificiales y técnicas de aprendizaje automático, el análisis clásico a disminuido pero no desaparecido[maj,].

La visión computacional clásica (análisis de complejidad algorítmica, optimización de recursos, etc) [Dawson-Howe, 2014] resulta importante al momento de explicar los formalismos detrás de las técnicas mas actuales. Por ejemplo, explicar una técnica que reside sobre una red neuronal profunda resultaría en una serie de tecnicismos sobre camadas, funciones de activación, probabilidad de *dropout* y muchas jerga obtusa; sin dar lugar a consideraciones directas : Que características extrae? , con que probabilidad selecciona cada area?, es eficiente en producción?.

Considerando todo esto, es claro que el Procesamiento digital de imágenes moderno es una mezcla de técnicas antiguas y nuevas, pero su aplcacion en la industria o servicios al publico es afín con el descongelamiento de la inteligencia artifical. Desde finales de los 80's, donde la venta de *scaners*con chips orientados al reconocimieto grafico de caracteres marcaba un comienzo en la industria del hardware orientado al procesamiento digital de imágenes y visión computacional. Este progreso continua hasta llegar a los actuales chips de IA y mejoramiento fotográfico embebidos en los celulares Huaweii o chips Nvidia Tegra [nvi,] .

La facilidad con la que el hardware a mejorado el desempeo de los algoritmos clásicos y los resultados de las técnicas modernas sobrepasaron o crearon nuevos enfoques, se ha conseguido alejarse del tele-procesamiento a sistemas embebidos mas robustos, por ejemplo el Curiosity Rover viajo equipado de un mini-clúster dual de chips PowerPC (elegidos por su muy bajo consumo energético) [Matthies et al., 2007].

II. Aplicaciones

II-A. Red Y.O.L.O.

En imágenes, You Only Look Once (YOLO) [Redmon et al., 2016] es un método de detección de objetos de enfoque avanzado. YOLO aplica una sola CNN a toda la imagen que divide la imagen en grillas. La predicción de los cuadros delimitadores y la puntuación de confianza respectiva se calculan para cada cuadrícula. Estos cuadros delimitadores se analizan según el puntaje de confianza previsto. La arquitectura de YOLO tiene 24 capas convolucionales y 2 capas completamente conectadas. La arquitectura se muestra en Figura 1. El algoritmo aplica una red neuronal a una imagen entera redimensionada a 480x480 píxeles. La red divide la imagen en una malla de SxS y presenta cuadros de límites, que son cuadros dibujados alrededor de imágenes y probabilidades para cada una de estas regiones. El método

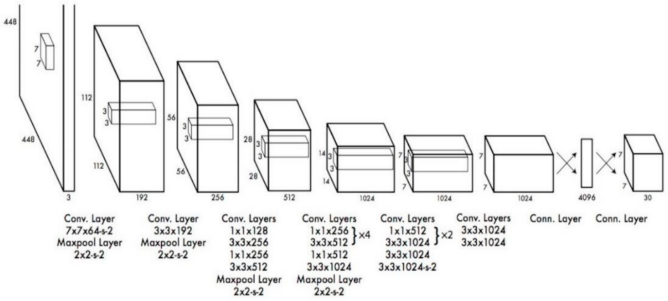

Figura 2. Arquitectura de YOLO.

Figura 3. Imagen de Entrada.

utilizado para generar estas probabilidades es la regresión logística. Los cuadros delimitadores están ponderados por las probabilidades asociadas. Para la predicción de clase, se utilizan clasificadores logísticos independientes. Sin embargo, la regla de un solo objeto limita lo cerca que pueden estar los objetos detectados. Para eso, YOLO tiene algunas limitaciones sobre lo cerca que pueden estar los objetos. En comparación con los métodos convencionales de detección de objetos, YOLO tiene ciertas ventajas. 1) En lugar de utilizar el método de dos pasos para la clasificación y localización del objeto, YOLO aplica un solo CNN tanto para la clasificación como para la localización del objeto. 2) YOLO puede procesar imágenes a aproximadamente 40-90 FPS, por lo que es bastante rápido. Esto significa que la transmisión de video se puede procesar en tiempo real, con una latencia insignificante en unos pocos milisegundos. La arquitectura de YOLO lo hace extremadamente rápido. Cuando se compara con R-CNN, es 1000 veces más rápido y 100 veces más rápido que el Faster R-CNN[Shinde et al., 2018].

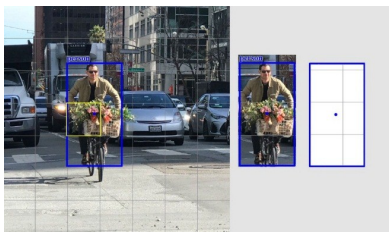

Figura 4. Imagen dividida en una malla SxS.

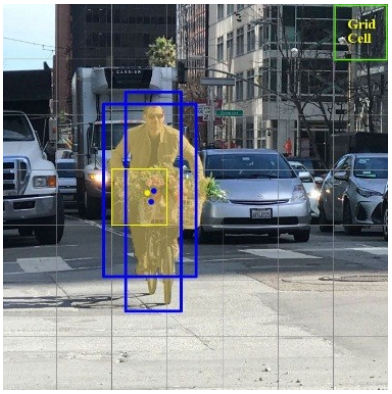

Figura 5. Imagen con cuadros delimitadores.

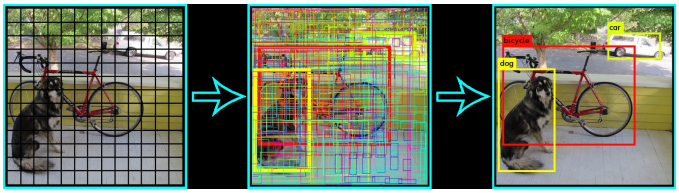

Figura 6. Entrada, Delimitadores, Objetos detectados.

II-B. Lentes de Realidad Virtual

?La Realidad Virtual es una simulación interactiva por computador desde el punto de vista del participante, en la cual se sustituye o se aumenta la información sensorial que recibe?. El hecho de que la simulación sea interactiva es lo que distingue la realidad virtual de una animación. En el caso de la simulación visual, los algoritmos de síntesis de imágenes calculan la imagen en el frame buffer con los parámetros de cámara virtual correspondientes a la posición y orientación de la cabeza del usuario. La información digital del frame buffer se traduce en una seal de video por la tarjeta gráfica y esta seal se traduce en una imagen física en el dispositivo de visualización.

Periféricos de Realidad Virtual

- Dispositivos de Entrada.
  - Posicionadores.
    - Magnéticos.
    - pticos
    - Acústicos.
    - Mecánicos.
    - De inercia.
  - Guantes de datos.
  - Registros de voz.
  - Dispositivos de entrada 3D.
- Dispositivos de Salida.
  - Efectos visuales.
    - Tecnologías de visualización.
    - Factores que determinan la calidad de visualización.
    - Dispositivos de visualización.
  - Efectos auditivos.

[Lee et al., 2019] Concluyen que la revolución tecnológica es un signo de nuestros tiempos. Las soluciones proporcionadas por la tecnología son aceptadas por la medicina y gradualmente introducidas en la práctica clínica diaria. Este artículo sirve como un resumen contemporáneo del entorno virtual y aumentado en el campo de la cirugía de tumores cerebrales.

II-C. Seguimiento Ocular (Eye Tracking)

El seguimiento ocular es la tecnología para resaltar el movimiento ocular. El dispositivo de implementación del seguimiento ocular es un rastreador de ojos. Las técnicas para obtener los datos del movimiento del ojo se pueden concebir en dos tipos: para medir el ojo posición relativa a la cabeza, y para medir la orientación de los ojos en el espacio, es decir, el ”punto de vista”[Duchowski, 2007].

El concepto de eye-tracking hace referencia a un conjunto de tecnologías que permiten monitorizar y registrar la forma en la que una persona mira una determinada escena o imagen, en concreto en qué áreas fija su atención, durante cuánto tiempo y qué orden sigue en su exploración visual.

La información que se extrae cuando se realiza la tecnología de Seguimiento Ocular,es la siguiente:

1. Dónde está mirando una persona de forma continua.
2. Qué le está llamando la atención y qué se la llamaba hace un momento.
3. Qué intenciones tiene esa persona.
4. El estado de ánimo de esa persona.
5. Donde debe ir colocado el contenido de valor para el cliente.
6. Si las seales visuales contenidas en la Web conducen de forma eficaz al cliente.
7. Capacidad del cliente para localizar la información que necesita en la Web.

Extraída esta información nos servirá para:

- Para mejorar la Estructura del Contenido.
- Mejorar la experiencia de los usuarios en la web.
- Guiar al usuario hacia el objetivo de su negocio.
- Facilitar los procesos que desea que realice el usuario.
- Conseguir Branding en Internet.
- Mejorar la imagen de marca a través del sitio web.

Se han hecho experimentos usando un rastreador ?Eyetribe? que tiene una precisión muy satisfactoria y puede producir buenos datos de salida, por lo que [Shokishalov and Wang, 2019] concluye que la práctica de seguimiento ocular proporciona información útil para el análisis e identificación de vulnerabilidades en los sistemas de gestión mediante documentos electrónicos.

II-D. The Azure Face API

La API de Azure Face es un servicio cognitivo que proporciona algoritmos para detectar, reconocer y analizar caras humanas en imágenes. La capacidad de procesar información de rostro humano es importante en muchos escenarios de software diferentes, como seguridad, interfaz de usuario natural, gestión y análisis de contenido de imágenes, aplicaciones móviles y

Figura 7. Mapa de calor, mide la intensidad de atencion.

robóticas. La Face API puede detectar rostros humanos en una imagen y devolver las coordenadas del rectángulo de sus ubicaciones. Opcionalmente, la detección facial puede extraer una serie de atributos relacionados con la cara, como postura de la cabeza, sexo, edad, emoción, vello facial y anteojos.

La API Verify realiza una autenticación contra dos caras detectadas o desde una cara detectada a un objeto de persona. En la práctica, evalúa si dos caras pertenecen a la misma persona. Esto es potencialmente útil en escenarios de seguridad. Para obtener más información, consulte la guía de conceptos de reconocimiento facial o la documentación de referencia Verificar API. La API Buscar similares toma una cara de destino y un conjunto de caras candidatas y encuentra un conjunto más pequeo de caras que se parecen más a la cara de destino. Se admiten dos modos de trabajo, matchPerson y matchFace. El modo matchPerson devuelve caras similares después de filtrar para la misma persona (usando la API Verify). El modo matchFace ignora el filtro de la misma persona y devuelve una lista de caras candidatas similares que pueden o no pertenecer a la misma persona.

II-E. Deep Fake

En los últimos 2 años con el avance en arquitecturas de redes neuronales profundas y el mejoramiento en técnicas de aumentado de muestras, se consiguió generar videos falsos a partir de pequeas muestras de video. Si bien este tipo de trabajos ya existían , cobro visibilidad con la presentación de un trabajo por la Universidad de Washington en el que se reconstruían expresiones faciales que pudieran acomodarse a un discurso especifico (en este caso, un discurso de Barack Obama) [Suwajanakorn et al., 2017].

Figura 8. Mapa de calor, mide la intensidad de atencion.

Sin embargo, trabajos mas recientes, ya no solo facilitan la sincronización de expresiones faciales o el modelado 3D apartir una imagen plana [Jackson et al., 2017], sino la creación de modelos 3D dinámicos [Berger et al., 2019].

Es importante considerar que todo esto es posible no tanto por las redes neuronales en si, sino por las t 'ecnicas de incremento artificial de datos, un área si bien relacionada a la vision computacional, es no dependiente.

En el futuro ***SIGGRAPH 2019*** dará debut mas de 4 papers asociado a técnicas de generación de video animación, creación de cuadros intermedio e iluminación dinámica. [you,]

III. Conclusiones

Uno de los principales motivos por el cual casi todas , por no generalizar, las aplicaciones y técnicas de visión computacional actuales dependen de las redes neuronales, recae en su *facilidad* de implementación (se construye la arquitectura pero no mas).

Consideremos la Red Y.O.L.O, las camadas de convolución aplican técnicas clásicas pero la magia sucede en las interconexiones entre neuronas, donde se ajustan pesos y otras variables asociadas la Red como un todo, convirtiendo una red de nodos que suman y restan valores en una función gigante con forma de grafo.

Mientras en las técnicas clásicas las funciones aplicadas a las imágenes (que no son mas que datos) debían ser escritas , demostradas y probadas de forma analítica; tal que el dominio y el rango siempre pasen por una función matemática formal. Lo que ocurre en las redes neuronales es que suplantan esta función analítica por una función paramétrica, y son estos parámetros los que permiten encontrar la mejor función que sea capaz de llevar del dominio a rango (ejm. imágenes a características).

Se remplazo la búsqueda *manual* de una función mágica capaz de cumplir con nuestro objetivo por una búsqueda *automática* (de ahí el nombre aprendizaje de maquina, pues es un proceso iterativo).

Finalmente, para estas ultimas técnicas, en las que los datos son vitales, no se debe olvidar la calidad de los mismos. No todos los datos son dominios y no todos los objetivos son rangos, esto debe ser tomado en consideración para evitar aplicar técnicas de avanzada en problemas sencillos.

Referencias

[nvi,] Introducing nvidia tegra 4, the world's fastest mobile processor.

[mit,] Mission and history, m.i.t.

[you,] Technical papers preview: Siggraph 2019.

[maj,] Visual computing market is likely to experience a tremendous growth in near future—nvidia corporation, intel corporation, advanced micro devices, arm.

[Berger et al., 2019] Berger, M., Li, J., and Levine, J. A. (2019). A generative model for volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 25(4):1636–1650.

[Dawson-Howe, 2014] Dawson-Howe, K. (2014). *A Practical Introduction to Computer Vision with OpenCV*. Wiley-IS&T Series in Imaging Science and Technology. Wiley, 1 edition.

[Duchowski, 2007] Duchowski, Z. (2007). Eye tracking techniques. *In Eye Tracking Methodology.*, pages 347–351.

[Jackson et al., 2017] Jackson, A. S., Bulat, A., Argyriou, V., and Tzimiropoulos, G. (2017). Large pose 3d face reconstruction from a single image via direct volumetric CNN regression. *CoRR*, abs/1703.07834.

[Keim et al., 2008] Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). *Visual Analytics: Definition, Process, and Challenges*, pages 154–175. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Lee et al., 2019] Lee, C., Wong, C., and Kwok, G. (2019). Virtual reality and augmented reality in the management of intracranial tumors. *Neuroscience.*, pages 14–20.

[Matthies et al., 2007] Matthies, L., Maimone, M., Johnson, A., Cheng, Y., Willson, R., Villalpando, C., Goldberg, S., Huertas, A., Stein, A., and Angelova, A. (2007). Computer vision on mars. *International Journal of Computer Vision*, 75(1):67–92.

[Redmon et al., 2016] Redmon, J., Girshick, R., Divvala, S., and Farhadi (2016). You only look once: Unified, real - time object detection. *IEEE Conference on computer vision and Pattern Recognition(CVPR).*, pages 779–788.

[Shinde et al., 2018] Shinde, S., Koyhari, A., and Vikram, G. (2018). Yolo based human action recognition and localization. *International conference on Robotics and Smart Manufacturing.*, pages 831–838.

[Shokishalov and Wang, 2019] Shokishalov, Z. and Wang, H. (2019). Applying eye tracking in information security. *13th International Symposium İntelligent Systems".*, pages 347–351.

[Suwajanakorn et al., 2017] Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13.