Texto completo del trabajo:

Seguridad en el entorno de la carretera a través de su entendimiento mediante visión por computador

Autores:

José María Armingol Moreno

Arturo de la Escalera Hueso

Abdulla Al Kaff

Fernando García Fernández

David Martín Gómez

** Todos ellos son investigadores en el Laboratorio de Sistemas Inteligentes de la Universidad Carlos III de Madrid*

Publicado como artículo en el número Especial de la Revista *Carreteras* 225 (julio/agosto 2019)

Título: Seguridad en el entorno de la carretera a través de su entendimiento mediante visión por computador

Introducción

En las últimas décadas, científicos e ingenieros, en la industria automovilística y de infraestructuras y en el ámbito docente e investigador, han desarrollado sistemas inteligentes de transportes para solucionar una gran variedad de problemas de seguridad en vehículos y sus infraestructuras viaria, adaptándose a las necesidades de la creciente sociedad segura en el ámbito de la automoción, el tráfico y las infraestructuras. Sin embargo, la digitalización de la carretera como infraestructura inteligente y los vehículos inteligentes que harán uso de dicha infraestructura son todavía dos retos muy importantes para los operadores de infraestructuras y fabricantes de vehículos, debido al coste actual de la tecnología necesaria para su implantación de forma eficiente y garantizar la seguridad completa. Por esta razón este trabajo plantea la seguridad del entorno de la carretera particularizado en el segundo reto actual, es decir, los vehículos digitales seguros a través del entendimiento del entorno mediante la visión por computador, los escáneres láser multiplano, los algoritmos inteligentes, las arquitecturas software para control y navegación, y las comunicaciones entre vehículos y las infraestructuras.

La seguridad es por tanto el objetivo principal del trabajo que se va a presentar en esta memoria. El último informe de presentado por la Dirección General de Tráfico a principios de 2018 recoge el balance anual de siniestralidad vial 2017 y resalta la necesidad de seguir trabajando en materia de seguridad para reducir la siniestralidad en las carreteas. En España durante el año 2017 se han producido 1.067 accidentes mortales en vías interurbanas, en los que han fallecido 1.200 personas y 4.837 heridas hospitalizadas, lo que supone un aumento del 3% en lo que a accidentes mortales (+28) y fallecidos (+39) se refiere y una disminución de un 6% (-336) en lo relativo a heridos hospitalizados.

Los sensores actuales de elevada precisión, que son diseñados para proporcionar una respuesta fiable frente a cambios inesperados en el entorno del vehículo requieren la interpretación de sus datos. Los volúmenes de datos que generan dichos sensores son de varios millones de mediadas por segundo, tanto de los sensores de visión por computador como los escáneres láser multiplano. Por lo tanto, se plantea en el ámbito de los vehículos inteligentes y/o autónomos la necesidad de entender dichos datos para traducirlos en seguridad mediante algoritmos inteligentes y arquitecturas software que trabajen en tiempo real para procesar el elevado volumen de datos generados por los sensores actuales en dichos vehículos inteligentes y/o autónomos. Este planteamiento se corresponde con el objetivo a conseguir: que es el aumento de la seguridad de los vehículos futuros a través del procesamiento y entendimiento de los datos procedentes de los millones de mediciones

precisas por segundo de los sensores de visión por computador y escáneres láser multiplano en tiempo real.

Los millones de mediciones por segundo del entorno inspeccionan el estado de la carretera, proporcionan entendimiento y facilitan la toma de decisión en tiempo real de conducción, de forma similar a un conductor humano que realiza una conducción manual. Aunque los resultados son similares –conducción por una carretera– la ventaja de los sistemas automáticos basados en dicha tecnología de visión artificial y láser, es el incremento del tiempo de respuesta de la unidad de control del vehículo inteligente o autónomo que puede preparar la activación de los sistemas de seguridad del vehículo en tiempo real antes de la situación de peligro. Es por esta razón, que ganar varios segundos de reacción, o incrementar el tiempo de respuesta de la maniobra automática del vehículo, es un gran avance con respecto a los vehículos actuales.

Por tanto, es necesario resaltar, que la alerta temprana de los sistemas automáticos de seguridad frente a situación de peligro se consigue mediante el entendimiento de los millones de mediciones por segundo del entorno. Esto es una realidad que será presentada y desarrollada en la presente memoria para mostrar los avances en el entendimiento seguro del entorno de carretera. Los algoritmos presentados en dicha memoria son una realidad de seguridad a través del entendimiento de millones de datos por segundo.

A continuación se resumen las investigaciones recientes en esta dirección de seguridad mediante el entendimiento del entorno de la carretera. Los vehículos que se han utilizado para el desarrollo de la tecnología de seguridad están basados en plataformas de vehículo inteligente y de vehículo autónomo, lo que le permite una gran versatilidad y abre un gran abanico de posibilidades a la investigación de algoritmos inteligentes y arquitecturas de percepción seguras (Figura 1).

Fig 1. Plataformas de investigación utilizadas en el entorno de la carretera

En el caso de los vehículos autónomos el hardware de las plataformas iCab (carritos de golf autónomos) están compuesto por cuatro bloques principales: El bloque actuador, el bloque de percepción, el bloque de comunicaciones y el bloque de procesamiento:

- Bloque actuador. Es el conjunto de sistemas electrónicos, diseñados para poder actuar sobre el vehículo autónomo. Consiste en un sistema de placas electrónicas, que permiten actuar sobre el motor eléctrico del vehículo y de un sistema de actuación en la dirección compuesto por un motor eléctrico, un sistema de engranajes y un encoder. Un actuador lineal permite actuar sobre el freno mecánico, dejando la posibilidad de éste de ser activado por un operado humano en caso necesario. Todos estos sistemas son desactivados por un sistema de parada de emergencia, que puede ser activado de forma remota, para las pruebas en modo autónomo.

- Bloque de percepción. Es el conjunto de sensores, tanto activos como pasivos, que permiten conocer tanto el estado del vehículo como el entorno que le rodea. Estos sensores son:

    - Cámara estéreo para la detección de obstáculos y localización en el entorno de la carretera
    - Lidar Velodyne para mapeado, localización y detección de obstáculos
    - Sistema de GPS y Brújula para posicionamiento auxiliar
    - Lidar SICK de un plano para detección de obstáculos, localización y mapeado del entorno
    - Sensores ultrasonidos para detección de obstáculos en campo cercano

- Bloque de comunicaciones. Permite la comunicación del vehículo, tanto con otros vehículos, como con diferentes agentes con los que interactúa. Para ello se ha incluido un router con tecnología 4G que permite en cada vehículo conectarse tanto a una red WiFi disponible, como a una red 4G de datos. Esto junto a una red virtual VPN permite diversas configuraciones de comunicaciones.

- Bloque de procesamiento. Los vehículos inteligentes y las plataformas incluyen ordenadores que realizan tareas de control y procesamiento de la información en tiempo real, especialmente diseñados para poder ser empleados en dispositivos embarcados. El control autónomo es una de las claves de la conducción autónoma, y por tanto se han diseñado numerosas soluciones para proveer a las plataformas, tanto de un control robusto a nivel bajo (acelerador, freno y dirección), como a nivel alto: cálculo de trayectorias, evitación de obstáculos mediante técnicas avanzadas de control inteligente basadas en el entendimiento del entorno, etc.

La comunicación entre vehículos, infraestructura y peatones forma parte de unos de los pilares de la seguridad, y dichas comunicaciones se han implantado con el objetivo de desarrollar vehículos autónomos y seguros. Por tanto, los vehículos desarrollados presentan sistemas de comunicaciones entre vehículos, basada en una red virtual privada (VPN) y arquitectura ROS, que permite la comunicación entre diversos agentes en la red, de forma descentralizada y escalable. Esto permite adaptarse a diferentes configuraciones de red, según las necesidades, además de un despliegue rápido que permite añadir nuevos vehículos sin comprometer la efectividad de la red.

A continuación, se muestra un ejemplo de la comunicación de la plataforma con la infraestructura, que se ha diseñado a través de una interfaz web, con el objetivo de conocer el estado de los vehículos además de realizar peticiones de servicio (Figura 2).

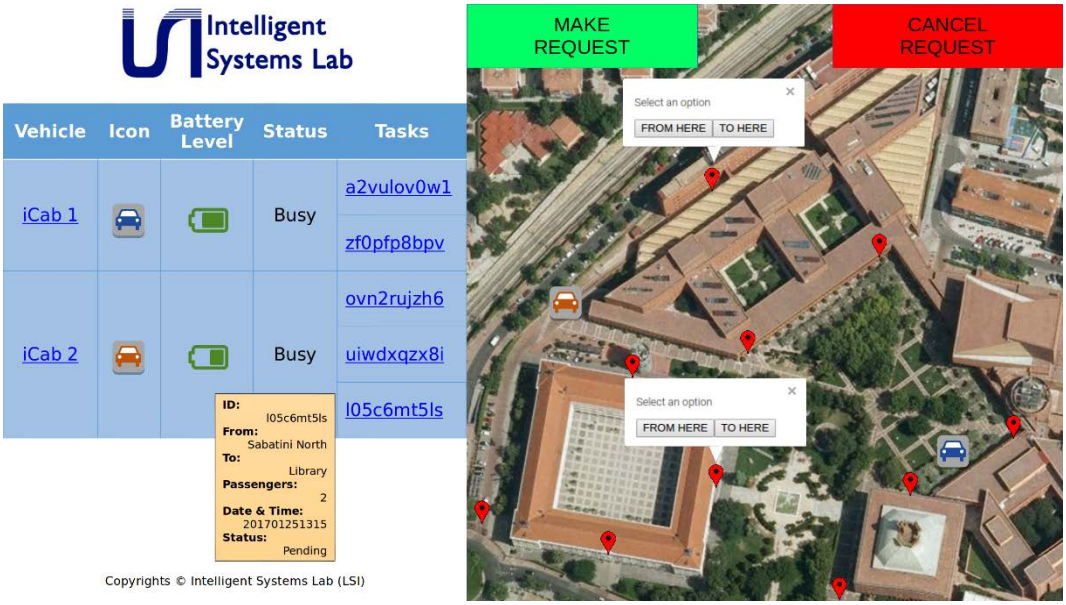


Fig 2. Interfaz web con comunicaciones integradas

Además, la comunicación entre un vehículo y un peatón también forma parte del concepto de seguridad en el entorno urbano, y por tanto, se ha diseñado una aplicación móvil, capaz de comunicarse con dispositivos móviles y de calcular una posible trayectoria de colisión, avisando del peligro tanto al usuario del móvil como al vehículo autónomo. La aplicación puede trabajar en segundo plano, avisando de la dirección en la que se acerca el vehículo autónomo (Figura 3).

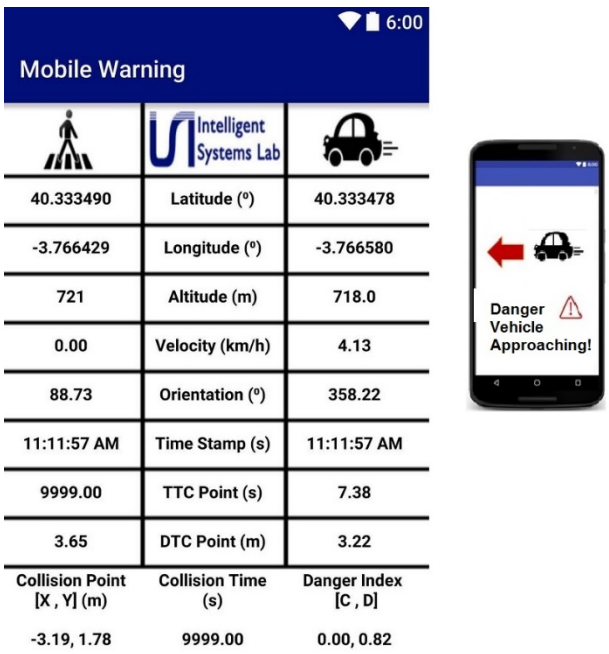


Fig 3. Ejemplo de la aplicación con el cálculo del tiempo de posición y el mensaje de aviso

Por otro lado, la cooperación entre vehículos también forma parte de la seguridad de los futuros vehículos, es decir, la realización de tareas de forma colaborativa basadas en técnicas como MRTA (Multiple Robot Task Allocation).

El objetivo de seguridad en vehículos inteligentes y autónomos por tanto está basado en el entendimiento del entorno para la navegación segura y requiere de técnicas inteligentes y procesamientos adaptados a la seguridad. Así uno de las técnicas que se pueden utilizar en la actualidad, es la fusión sensorial para la detección de obstáculos, donde los diferentes sistemas de percepción incluidos en el vehículo se fusionan para permitir una detección de obstáculos y un diseño de mapas avanzado y robusto. Un ejemplo se muestra en la Figura 4.

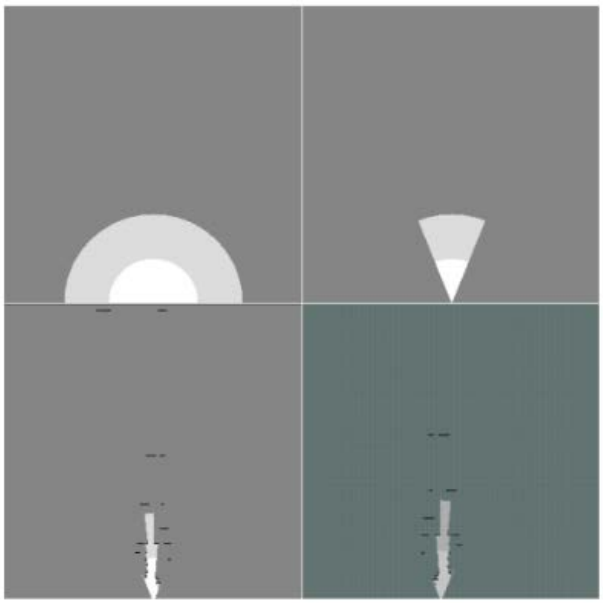


Fig 4. Ejemplo del mapa de fusión sensorial (abajo) con un escáner laser (izquierda) y una cámara (derecha)

Sin embargo, son las técnicas de visión por computador y láser multiplano, las que están dando mejores resultados de seguridad por su alta eficiencia y capacidad para entender el entorno del vehículo. Por ejemplo, las técnicas de clasificación semántica a partir de procesamiento de imágenes por computador, permite la clasificación del entorno, así como, la detección de obstáculos en la vía. Estos algoritmos de visión por computador han permitido desarrollar una tecnología que es capaz de identificar cada elemento de una imagen, tanto su información tridimensional (obstáculo o camino libre), como clasificarlo mediante técnicas de inteligentica artificial (Figura 5).

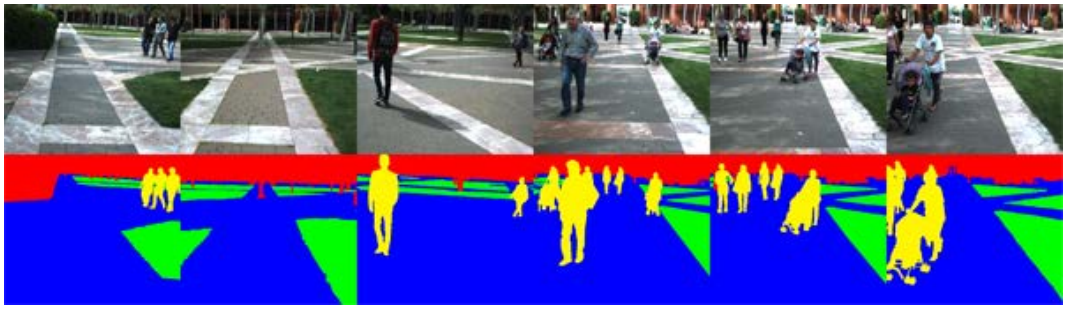


Fig 5. Ejemplos de clasificación semántica, camino transitable (Azul), peatones (amarillo), zona no transitable (verde)

Otra de las técnicas de visión por computador actuales es la odometría visual, que nos permite conocer el avance en una carretera donde los sistemas de posicionamiento tradicionales como el GPS no funcionan correctamente. Es decir, mediante tecnología estéreo es posible conocer información tridimensional, además de la información visual. Esto permite que sea posible reconstruir el movimiento del propio vehículo en un entorno, permitiendo dotar de capacidades de localización a un vehículo utilizando únicamente visión por computador (Figura 6).

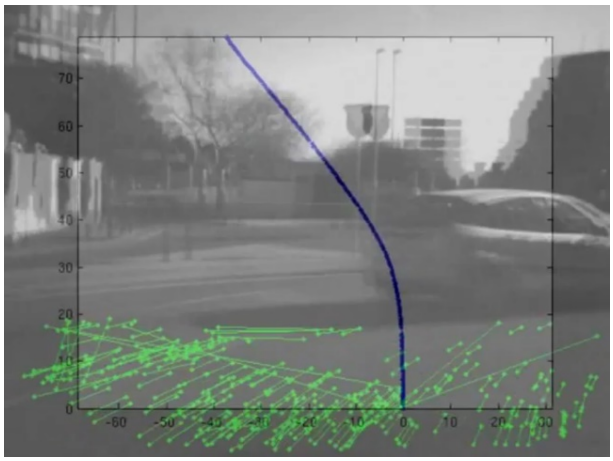


Fig 6. Ejemplo de la reconstrucción trayectorias y posicionamiento utilizando odometría visual

La tecnología láser multiplano, también está permitiendo grandes avances en el entendimiento de los obstáculos del entorno de la carretera, como es el mapeado 3D. Esta técnica además permite una localización segura en un entorno estructurado, donde los diferentes sensores disponibles en el vehículo permiten realizar un mapa 3D de las localizaciones. El resultado se muestra en la Figura 7.

Fig 7. Reconstrucción del entorno de navegación mediante láser multiplano

El láser multiplano además permite aplicar las conocidas técnicas de localización y mapeado simultáneo (SLAM), es decir, permiten un mapeado y localización del vehículo autónomo basado en los diferentes sensores incluidos en el vehículo (Figura 8).

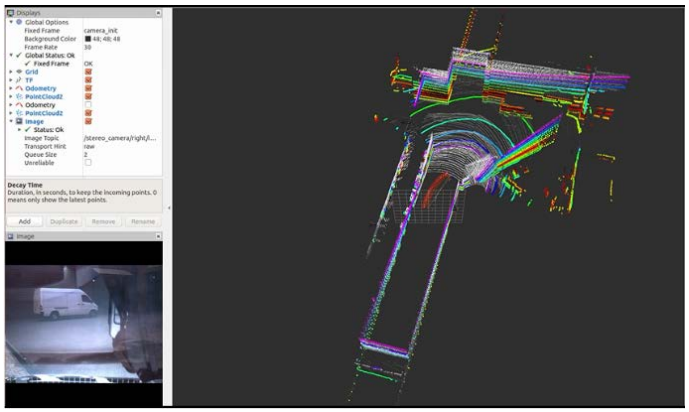


Fig 8. Ejemplo de SLAM mediante Lidar

Por estas razones, se va a presentar en este trabajo, los últimos avances en materia de seguridad a partir de métodos de visión por computador, que nos permiten entender correctamente el entorno de la carretera. A continuación, en los siguientes apartados, se detallan las tecnologías desarrolladas para el avance en materia de seguridad de vehículos en la carretera mediante visión por computador. En resumen, esta memoria, está basada en un primer apartado con la introducción a las tecnologías que estamos desarrollando en la actualidad, y un segundo apartado donde se especifica nuestra investigación puntera relativa a seguridad en el entorno de la carretera mediante visión por computador

2. Modelado de escenas de tráfico para vehículos inteligentes en carretera utilizando estimación de detección y orientación basada en una Red Neuronal Convolucional

La tecnología de seguridad en el entorno de la carretera ha adoptado un papel cada vez más importante en los sistemas de transporte en las últimas décadas. Los sistemas avanzados de asistencia al conductor (ADAS) han sido desarrollados para lidiar con la toma de decisiones equivocada y las distracciones de los humanos que son la principal causa de los accidentes de tráfico. Estos sistemas "*de seguridad inteligente*" representan un aumento en el grado de automatización de la seguridad hacia el objetivo futuro de la conducción totalmente autónoma. En este momento existen vehículos autónomos que han sido probados con éxito en los entornos urbanos [1], sin embargo en la mayoría de los casos son muy dependientes de los mapas digitales.

La navegación autónoma con conocimiento previo escaso o inexistente del entorno sigue siendo un desafío para la seguridad debido a la amplia gama de situaciones complejas que pueden aparecer (puntos de referencia ocluidos, comportamientos inesperados, etc.) dentro de un entorno altamente dinámico y semiestructurado. Por esta razón los futuros sistemas de conducción automática basados en percepción completa del entorno serán necesarios para comprender las complejas situaciones del tráfico en una carretera por ellos mismos. Este requisito se conseguirá mediante un entendimiento pleno y automático de la posición y el movimiento de cada vehículo en la carretera.

En la carretera, no solo debe tenerse en cuenta la presencia de los obstáculos en frente del vehículo, sino también una estimación precisa de la clase de obstáculo (es decir, automóvil, ciclista, etc.), para comprender y predecir correctamente la situación del tráfico y el estado de la carretera.

Los enfoques basados en visión por computador [2] han demostrado ser muy útiles para su uso en vehículos inteligentes por la eficaz percepción del entorno y por su tamaño compacto y facilidad de integración. Por esta razón, el presente trabajo se centra en la seguridad mediante sensores de visión por computador. Esto es, un enfoque basado en visión por computador que permite entender el entorno de la carretera.

A continuación se va a presentar nuestros trabajos dirigidos a la detección y localización de los elementos presentes en una carretera. Además, la detección de objetos de la carretera se va a enriquecer a través de una estimación del punto de vista, permitiendo inferencias de alto nivel sobre comportamientos a corto plazo. El algoritmo estará basado en una Red Neuronal

Convolucional (CNN) moderna para realizar la inferencia crítica de acuerdo con las características de la carretera. Además, la información estéreo de la cámara binocular permitirá el razonamiento espacial en el sistema inteligente del vehículo.

2.1 Estado del arte

La detección de obstáculos es una característica esencial para los sistemas de conducción automatizados. Por esta razón, una gran cantidad de algoritmos se han desarrollado históricamente con este propósito. El esfuerzo a menudo se centra en la detección de vehículos y peatones ya que estos obstáculos son los más comúnmente encontrados en escenas de tráfico y carretera.

De acuerdo con el dispositivo de detección en uso, los métodos basados en visión por computador se pueden dividir en dos categorías principales: métodos de visión monocular y métodos de visión estéreo. La visión estereoscópica proporciona información de profundidad sobre la escena y por lo tanto es comúnmente utilizada en aplicaciones de conducción [3]. Los algoritmos de visión estereoscópica generalmente hacen suposiciones sobre el terreno o el espacio libre esperado en la carretera[4]. Sin embargo, la información de la geometría la escena puede ser recuperada, permitiendo la construcción de representaciones tales como mapas de ocupación probabilísticos [5], mapas de elevación [6] o modelos 3D completos [7], donde los obstáculos pueden ser identificados. Por otro lado, la detección monocular de obstáculos se basa comúnmente en características de apariencia. La selección de características adecuadas ha sido tradicionalmente la parte más importante de los sistemas de visión por computador, es decir, es una etapa crucial en el flujo de ejecución y rendimiento de un algoritmo de visión por computador, y se han desarrollado numerosas aplicaciones con características específicas, como HOG-DPM [8], para detectar los usuarios de la carretera (por ejemplo, ciclistas [9]). La estimación de orientación de la los usuarios o elementos de la vía detectados, son menos frecuentes, pero también se ha estudiado por varios autores [10].

La representación del aprendizaje mediante redes neuronales profundas ha conducido a un cambio de paradigma en los últimos años, mostrando grandes mejoras con respecto a los métodos de extracción de características de forma manual para posteriores tareas de reconocimiento. En particular, es este trabajo utilizaremos las Redes Neuronales Convolucionales (CNN), porque pueden aprender representaciones de datos jerárquicos y se ha demostrado su alta utilidad en la clasificación de objetos en un entorno de carretera [11].

En lugar de utilizar el enfoque clásico de ventana deslizante, la detección con CNN se va a basar en mecanismos de atención para limitar el número de propuestas para ser clasificadas eficientemente.

Dentro de esta tendencia, Girshick et al. introdujo el método de reconocimiento usando regiones (R-CNN) [12]. Estas regiones se pueden seleccionar según los métodos de segmentación basados en la similitud clásica de ventana deslizante; sin embargo, se ha hecho un gran esfuerzo recientemente en los flujos de datos de la red, donde cada etapa se puede aprender de manera efectiva. En este sentido, el método rápido R-CNN [13], aprovecha una Red de Propuestas de Región (RPN) que alimenta a la R-CNN que es la responsable de la tarea de clasificación.

En la actualidad, se han aplicado CNNs en varias tareas relacionadas con la conducción autónoma, como la detección de los carriles de la carretera [14] y, por supuesto, la detección de vehículos y obstáculos en la carretera [15]. En algunos casos, la orientación también se puede predecir para incrementar la información de las detecciones; por lo tanto, en [16], se usa una CNN para la detección de objetos y la estimación de los ángulos de orientación. Los ángulos de orientación se estiman en [17] mediante modelos de probabilidad.

2.2 Resumen del sistema

Este trabajo ha sido desarrollado para utilizarse en los vehículos inteligentes y autónomos presentados previamente, donde dicho trabajo se ha empezado a probar en el vehículo IVVI 2.0 (Vehículo inteligente basado en información visual) [18]. El IVVI 2.0 es un vehículo inteligente que es conducido de forma manual pero está equipado con tecnología de visión por computador necesaria para probar dichos algoritmos de entendimiento del entorno de la carretera para navegación segura.

El sistema de visión por computador incluye una cámara estéreo trinocular, que proporciona las imágenes del entorno de la carretera que se han utilizado en el presente trabajo. La unidad de procesamiento embebida en el coche incluye una GPU de alto rendimiento que permite el procesamiento en paralelo, como el procesamiento llevado a cabo en la red neural convolucional (CNN). Además, la arquitectura Robot Operative System (ROS) se utiliza para la cooperación entre módulos de procesamiento de datos.

El método presentado en esta memoria proporciona por tanto un paso adelante en la detección y clasificación basada en visión por computador para el entendimiento del entorno de la carretera. En resumen, el trabajo presentado consiste en dos ramas principales que están diseñadas para funcionar en paralelo, como se muestra en la Fig. 1:

1. Detección de objetos y estimación de ángulos basado en la apariencia del entorno de la carretera. Las características se extraen exclusivamente de la imagen estéreo izquierda.

2. Localización de los objetos de la carretera, mediante el desarrollo de un método robusto frente a los cambios de posición y orientación del sistema de visión por computador debido al movimiento del vehículo. El método está basado en una reconstrucción 3D utilizando visión estéreo, donde los parámetros extrínsecos de la cámara son extraídos considerando el plano el suelo de la carretera.

Como es habitual en los métodos de aprendizaje profundo, como es nuestro caso en la detección de objetos en la carretera, el procesado del algoritmo se realizará completamente en una GPU. Por otro lado, en la parte del algoritmo relativa a la localización del objeto en la carretera, se realizará un uso intensivo de CPU en la etapa de reconstrucción 3D. Este doble proceso ha sido diseñado para optimizar y maximizar la capacidad de procesamiento disponible, con el fin de cumplir con los requisitos de tiempo inherentes a la aplicación de visión por computador.

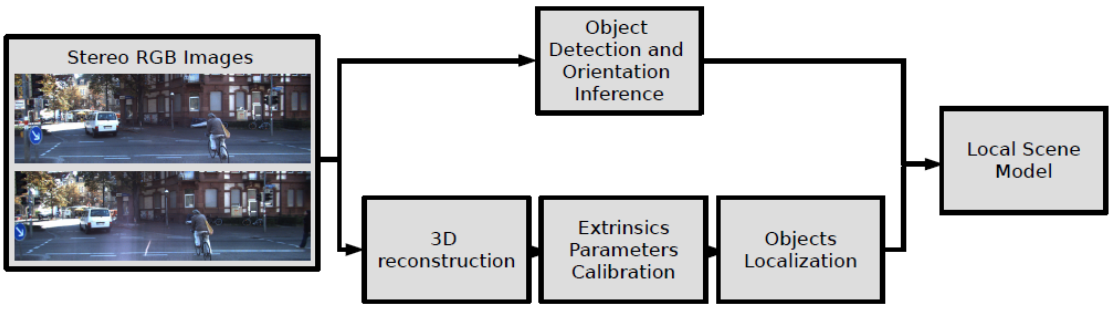


Fig. 9. Visión general del sistema de visión por computador para entendimiento seguro de la carretera propuesto

2.3 Detección de obstáculos

En el entorno de la carretera se puede encontrar una gran variedad de obstáculos dinámicos, desde el entorno urbano hasta el entorno de carretera secundaria o autopista. Las técnicas de clasificación de objetos permiten clasificar regiones de imágenes predefinidas, sin embargo, se ha demostrado que en el entorno de la carretera, por seguridad, también es necesario localizar cada objeto en la imagen con sus correspondientes coordenadas.

En esta memoria se adopta por tanto un enfoque nuevo basado en las redes rápidas, llamadas R-CNN rápidas [13], para realizar la detección de objetos. Por tanto, basándonos en el detector R-CNN [12], una red R-CNN rápida proporciona un marco para entrenamiento que abarca desde los píxeles de la imagen hasta la predicción final.

Por tanto mientras se van mejorando los algoritmos de detección clásicos, la técnica R-CNN rápida puede encargarse de procesar un gran número de clases sin disminuir el rendimiento, y es por esta razón, que la técnica propuesta es muy apropiada para los entornos de carretera.

La técnica R-CNN rápida está basada en dos etapas: una red de propuestas de regiones del entorno de la carretera (RPN), que es responsable de identificar las regiones de la imagen donde los objetos están ubicados, y una red R-CNN, donde las regiones de la imagen provenientes de la red anterior RPN son clasificadas. De esta forma, las dos etapas, se basan en arquitecturas de redes neuronales convolucionales (CNN), y de hecho, ambas etapas comparten el mismo conjunto de capas convolucionales. Así, la red R-CNN rápida permite la detección de objetos en tiempo real, y por tanto, consigue el necesario aumento de seguridad en el entorno de la carretera.

En este trabajo además, se ha adoptado la estrategia introducida en [19] para incorporar la inferencia del ángulo en el marco de la detección. Dicha estrategia permite mejorar el entendimiento del entorno de la carretera mediante visión por computador.
La base de la idea es beneficiarse de las características convolucionales ya calculadas para obtener una estimación de la orientación de los objetos con respecto a la cámara. La Figura 10 ilustra el método propuesto en detalle. Así, como con las propuestas de regiones en la red RPN, en este caso el ángulo se puede estimar casi sin coste computacional durante el tiempo de prueba dado que las convoluciones se calculan solo una vez.

De acuerdo con los requisitos de la aplicación, solo se estima el ángulo de guiñada (es decir, el acimut), que es el ángulo desde el que se ven los objetos. Esto es debido a que los obstáculos en el entorno de la carretera y el movimiento propio del vehículo, se estima que se mueven en el mismo plano de la carretera.

2.3.1 Estimación discreta del ángulo de los obstáculos en el entorno de la carretera

La solución propuesta adopta un enfoque discreto para la estimación del ángulo de los objetos presentes en la carretera, así el rango completo de los ángulos posibles (2π rad) se consigue dividirlo en N_b medidas θ_i; es decir, $i = 0, \dots, N_b - 1$, de los cuales solo una medida se emplea para representar el ángulo del objeto u obstáculo en el entorno de la carretera.

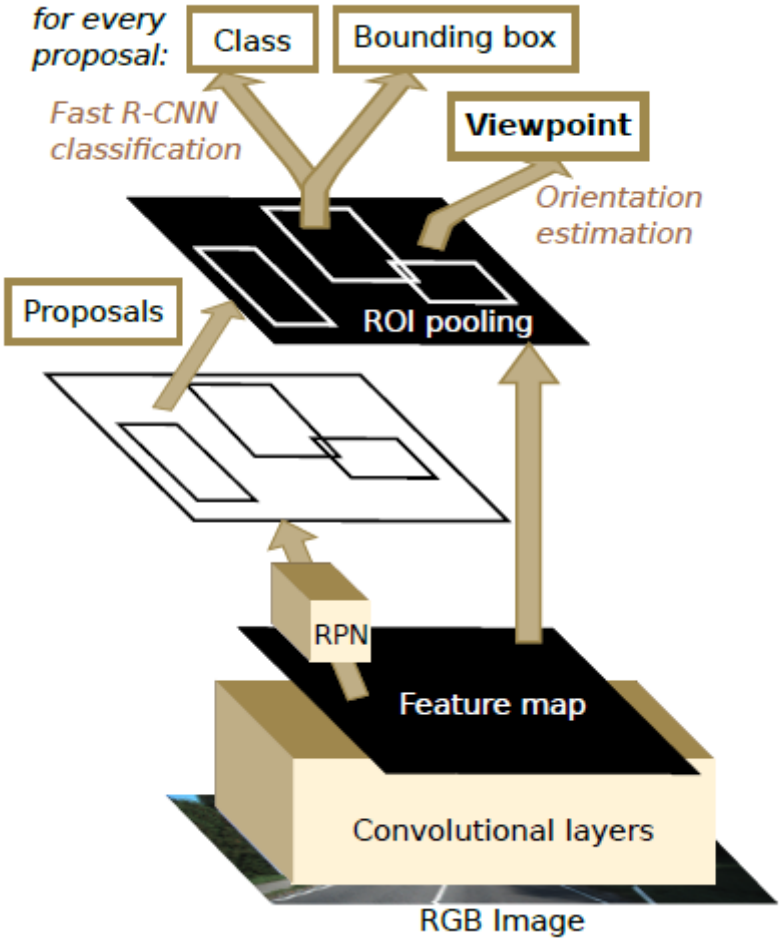


Fig. 10. Propuesta de detección de objetos y enfoque de estimación de puntos de vista.

Por tanto, los obstáculos de la carretera con un ángulo θ calculado manualmente como referencia correcta, son asignados con una etiqueta *i* durante la etapa de entrenamiento, de tal forma que $\theta \epsilon \Theta_i$:

$$\Theta_i = \left\{ \theta \in [0, 2\pi) \;\middle|\; \frac{2\pi}{N_b} \cdot i \leq \theta < \frac{2\pi}{N_b} \cdot (i+1) \right\} \qquad (1)$$

La estimación propuesta del ángulo del objeto de la carretera tiene como objetivo proporcionar un ángulo estimado que consiste en una distribución categórica de parámetros sobre N_b posibles ángulos, r. Una estimación única $\hat{\theta}$ puede por lo tanto ser proporcionada como el centro de la medida b* con la mayor probabilidad según r:

$$\hat{\theta} = \frac{\pi(2b^* + 1)}{N_b} \tag{2}$$

2.3.2 Estimación del ángulo y de la detección conjunta

En el marco la red R-CNN, las regiones de la imagen se propagan a través de la red y, finalmente, se extrae un vector de características de longitud fija para predecir la clase de objeto, es decir, para predecir el objeto que está delante del vehículo en la carretera.

Por tanto, en nuestro trabajo introducimos la estimación del ángulo de forma directa. Dicho ángulo es inferido desde el mismo vector de características que se usa también para predecir la clase del objeto de la carretera. Esto está motivado por el hecho de que la apariencia está muy influenciada por el ángulo del objeto de la carretera. De esta forma, un buen conjunto de características debería ser capaz de discriminar entre diferentes ángulos.

Las soluciones obtenidas con el método R-CNN rápido [20] se utilizan en este trabajo, por lo que los vectores de características resultantes son introducidos en una secuencia de capas totalmente conectadas que son finalmente divididas en tres capas hermanas. Como en el enfoque original del método, los dos primeras capas hermanas proporcionan una clasificación y una regresión de límites del objeto, respectivamente.

Por otro lado, la nueva tercera capa es responsable de dar una estimación del ángulo, que en última instancia se normaliza a través de una función softmax. Dado que la clasificación se realiza sobre K clases, la salida de esta rama es un vector r compuesto de $N_b \cdot K$ elementos, representando K distribuciones categóricas (una por clase) sobre las medidas de ángulo N_b:

$$\boldsymbol{r}^k = (r_0^k, \ldots, r_{N_b}^k) \text{ for } k = 0, \ldots, K \tag{3}$$

2.3.3 Detalles de implementación del método

Como es habitual en las tareas de clasificación, se espera que las capas convolucionales se inicialicen utilizando previamente un modelo entrenado con el conjunto de datos de clasificación de ImageNet [21], mientras que las capas completamente conectadas reciben valores aleatorios de acuerdo con una distribución Gaussiana. En este trabajo, se consideran ocho medidas de ángulo uniformemente espaciadas para la estimación del ángulo (Nb = 8).

Finalmente, se aplica un método de supresión no máxima por clase (NMS) para eliminar las detecciones duplicadas.

4. Modelado de la escena de la carretera

La detección de objetos se puede mejorar con información geométrica para recuperar un modelo instantáneo y local de los objetos presentes en la carretera frente al vehículo. Para lograr este objetivo, utilizamos la información de las dos cámaras pertenecientes al sistema de visión estéreo para obtener una reconstrucción densa 3D del entorno que está frente al vehículo, es decir, la carretera y sus obstáculos.

Inicialmente, la nube de puntos 3D de la escena se representa en las coordenadas de la cámara. Si se supone que el suelo es plano en una pequeña vecindad en frente del vehículo, entonces los parámetros extrínsecos del sistema de visión estéreo pueden ser estimados fácilmente. En consecuencia, el efecto de los cambios de la posición y orientación de la cámara debido al movimiento del vehículo (por ejemplo, viajar en superficies irregulares de la carretera) puede eliminarse correctamente.

A través de este proceso, los obstáculos detectados en la etapa de detección de objetos de la carretera, se pueden localizar en coordenadas del mundo y asignarles un ángulo de orientación absoluto.

4.1 Reconstrucción con sistema de visión estéreo 3D

En este trabajo adoptamos la técnica “semi-global” [22] para realizar el emparejamiento denso de puntos del sistema estéreo, es decir, la correspondencia entre los puntos detectados en la imagen izquierda con los puntos detectados en la imagen derecha.
A pesar de que esta familia de algoritmos necesita más procesamiento que los métodos tradicionales de comparación de bloques, los retos planteados por el entorno de la carretera (como por ejemplo la falta de textura o los cambios de iluminación), hacen que sean los más adecuados para el objetivo de este trabajo. Como ejemplo, el mapa de disparidad obtenido de la escena correspondiente a la Figura 11a se muestra en la Figura 11b.

Como resultado, se obtiene una nube de puntos tridimensional (Figura 11c). Que está compuesta por una rejilla de cubos de menor resolución, con un tamaño de cuadrícula de 20 cm. Este método, además reducir la cantidad de datos para ser procesados, dicho filtrado tiene el objetivo de normalizar la densidad de puntos a lo largo del eje de profundidad.

4.2 Calibración automática de los parámetros extrínsecos

Los coeficientes del plano de la carretera se deben estimar como un primer paso para obtener los parámetros extrínsecos del sistema de visión. En dicho método se aplican dos filtros de pasa banda para eliminar los puntos fuera de un rango de [0 2] m a lo largo del eje vertical y un rango de [0 20] m a lo largo del eje de profundidad. Dentro de esos rangos, la suposición de suelo plano se cumple con alta probabilidad

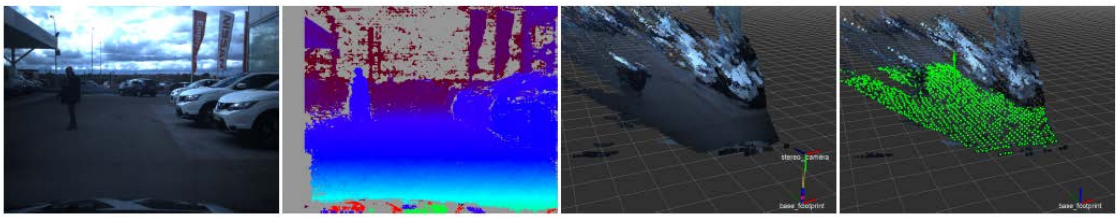
Los puntos que comprenden la nube de puntos filtrada se adaptan a un plano usando RANSAC [23] con un umbral de 10 cm. Además, solo se consideran los planos perpendiculares a una dirección fija, con una pequeña tolerancia angular. Debido a que los ángulos que definen la posición y orientación de la cámara son muy pequeños, dicho eje es elegido como la dirección vertical en coordenadas de la cámara. La Figura 11d presenta el plano de la carretera (mostrado en verde) que es obtenido de la nube de puntos de rejilla calcula previamente.

A continuación, se puede demostrar [24] que, dado un plano de la carretera definido por $ax_c + by_c + cz_c + d = 0, con\ (x_c, y_c, z_c)$ siendo las coordenadas de un punto que pertenece al plano de la carretera, entonces el balanceo (ψ), el cabeceo (ϕ) y la altura (h) que definen la posición y orientación de la cámara se pueden obtener como:

$$\psi = \arcsin(a) \qquad \phi = \arctan\left(\frac{-c}{b}\right) \qquad h = d \tag{5}$$

El ángulo de guiñada u orientación no se puede extraer únicamente del plano de la carretera, y por lo tanto se supone ser nulo. Además, en este trabajo se ha seleccionado no traducir las coordenadas del mundo a lo largo de los ejes x e y de la cámara, aunque ese desplazamiento puede elegirse arbitrariamente (por ejemplo, el origen puede estar centrado en la parte delantera del vehículo).

Ese conjunto de parámetros extrínsecos define una transformación que luego se aplica a la nube de puntos no filtrada para obtener los puntos en coordenadas del mundo.



(a) (b) (c) (d)

Fig. 11. Estimación de los parámetros extrínsecos: (a) imagen izquierda; (b) mapa de disparidad; (c) nube de puntos; (d) puntos pertenecientes al plano de la carretera, en verde, sobre la nube de rejilla.

4.3 Localización de objetos en el entorno de la carretera

En este apartado se explica que para obtener la ubicación espacial de los objetos en la escena, se utiliza la correspondencia entre puntos en la imagen y los puntos en la nube 3D, conservados dentro de una estructura de nube de puntos organizada. Los puntos que pertenecen al suelo de la carretera, así como aquellos que están demasiado cerca de la cámara (por ejemplo, los correspondientes al capó del vehículo), se eliminan inicialmente. Así, para cada detección, los valores medios de las coordenadas (x,y,z) para el conjunto de los puntos 3D correspondientes a las 11 filas centrales de la caja contenedora del objeto, son calculados y usados como una estimación de la localización 3D del objeto. El ángulo de guiñada u orientación, que es expresado como la rotación alrededor de un eje vertical local, puede aproximarse teniendo en cuenta el ángulo entre el eje x positivo de la coordenada del mundo y el punto dado por las coordenadas del objeto.

Al usar toda la información inferida sobre los obstáculos, se construye un modelo de vista de pájaro del entorno del vehículo, donde cada objeto en el campo de visión es incluido junto con su orientación estimada.

5. Resultados

El modelo conjunto propuesto de detección de objetos en la carretera y estimación de ángulo ha sido cuantitativamente evaluado de acuerdo con las métricas estándar en un conjunto de imágenes de referencia bien establecidas en la comunidad científica, mientras que el rendimiento de la etapa de modelado de escena y, finalmente, el sistema completo, han sido probados en escenas reales de tráfico usando los vehículos presentados al inicio de esta memoria.

5.1 Detección de objetos y estimación de ángulos

En este apartado se presenta en primer lugar los experimentos para evaluar la detección de objetos y la estimación de ángulos en el entorno de la carretera, llevado a cabo utilizando la base de datos de detección de objetos del KITTI [25], que proporciona las etiquetas de orientación y clasificación de cada imagen. Dado que las anotaciones para el conjunto de prueba no están públicamente disponibles, el conjunto de entrenamiento etiquetado se ha dividido en dos grupos, para entrenamiento y para validación, asegurando que las imágenes de la misma la secuencia no se usa en ambos subconjuntos. En total se han utilizado 5.576 en el entrenamiento, mientras que 2.065 imágenes se han empleado posteriormente para probar nuestros algoritmos.

Dado que nuestro trabajo se centra en el reconocimiento simultáneo de los diferentes agentes de la escena (coches, motos, peatones, etc), nuestro algoritmo ha sido entrenado para detectar las siete clases proporcionadas por el conjunto de datos del KITTI. Además, se tuvo especial cuidado para evitar incluir regiones que se superpusieran con las regiones DontCare y Misc, ni muestras positivas ni negativas durante el entrenamiento.

Dado que nuestro enfoque es independiente de la arquitectura particular seleccionada para las capas convolucionales, hemos probado las dos arquitecturas base de R-CNN rápida en nuestra aplicación: ZF [26] y VGG16 [27]. Por otro lado, aunque todos los modelos se obtuvieron escalando las imágenes de entrada a 500 píxeles en altura durante el entrenamiento, diferentes escalas han sido evaluadas en la fase de prueba. En todos los casos, el entrenamiento se llevó a cabo para 90.000 iteraciones, con una tasa de aprendizaje base de 0.0005, que fue escalada en 0.1 cada 30.000 iteraciones.

Con el objetivo de mostrar los resultados claramente, solo evaluamos el promedio de similitud de orientación (AOS), como se presenta en [25], que está destinado a evaluar conjuntamente la precisión de la detección y el ángulo de orientación. Los resultados obtenidos para las diferentes combinaciones de arquitectura y escala se muestran en la Tabla 1. En la tabla no aparecen los resultados de las clases “Persona Sentada” y “Tranvía” porque no son fiables debido al reducido número de muestras. Los tiempos de procesamiento se corresponden con la implementación utilizando la interfaz en Python de Caffe [28] y una tarjeta GPU Titan Xp de NVIDIA.

Tabla 1. Promedio de similitud de orientación (%) y tiempos de ejecución (ms) en la prueba para diferentes escalas y arquitecturas

Net	Scale	Car	Pedest.	Cyclist	Van	Truck	mean	Time
ZF	375	44.2	35.6	16.1	8.5	3.2	21.5	46
	500	52.7	43.7	18.4	12.9	3.5	26.2	73
	625	51.6	40.7	22.7	15.1	5.3	27.1	90
VGG	375	64.8	54.7	25.0	22.9	8.5	35.2	79
	500	74.7	61.0	33.0	30.0	12.1	42.2	112
	625	75.7	60.9	35.2	31.1	15.4	43.7	144

Como se muestra, la precisión no crece significativamente cuando la escala de tiempo de prueba es elevada, más allá de la escala de tiempo de entrenamiento, es decir, 500 píxeles. Por otra parte, VGG16 supera considerablemente a ZF para cada clase analizada.

5.2 Modelado de la escena

Las pruebas para el modelado de escena se realizaron utilizando los vehículos presentados inicialmente (Fig. 1) en situaciones reales de tráfico. De acuerdo con los resultados en la sección anterior, elegimos la arquitectura VGG16 y 500 píxeles como la escala de la imagen. Debido a la capacidad de generalización presentada por las estructuras de las redes neuronales convolucionales (CNN), los modelos entrenados con el conjunto de datos KITTI fueron utilizados sin modificaciones. Una región de interés (ROI) con 500 píxeles de altura, que comprende el área donde los objetos están típicamente presentes en las imágenes, se ha extraído a partir de la imágenes originales de 1024x768 píxeles para ser utilizadas por una rama de la red CNN, mientras que la imagen completa se emplea para construir la nube de puntos en la rama de modelado.

La Figura 12 muestra cuatro ejemplos de detecciones monoculares (fila superior) y sus modelos de escena resultantes, donde los obstáculos se representan como puntos en una vista de pájaro de la nube de puntos reconstruida (fila inferior). La orientación del objeto es representada por una flecha. Además, los puntos que pertenecen al plano de la carretera se proyectan en la imagen y se pintan en verde. Dichos puntos, proporcionan una estimación aproximada del área transitable para el vehículo en la carretera.

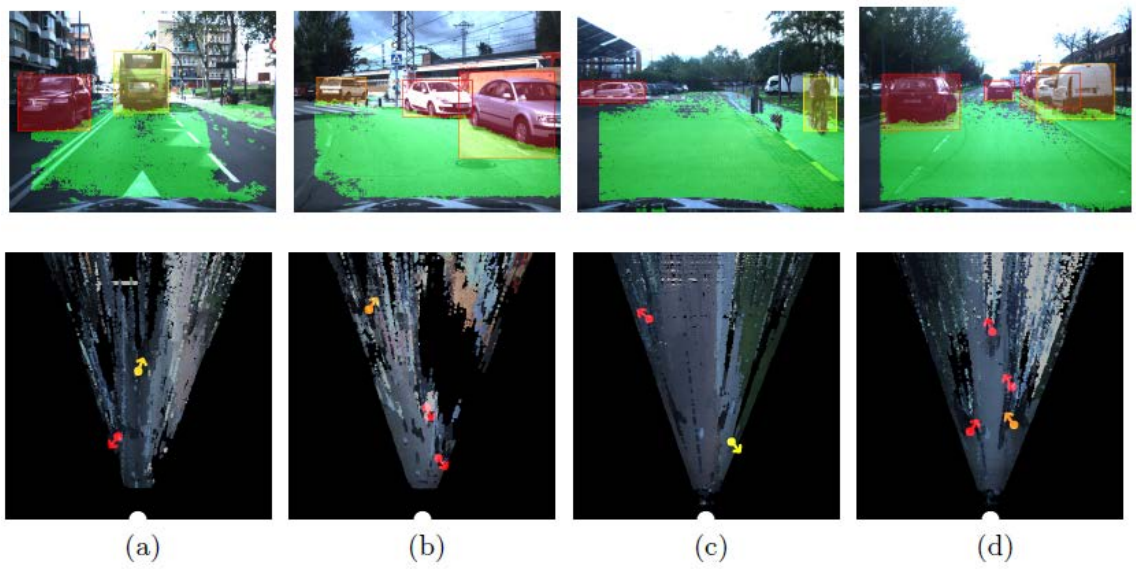


Fig. 12. Algunos ejemplos de escenas de tráfico real correctamente identificadas por nuestro sistema de visión por computador

Finalmente, resaltar que la continuidad de este trabajo ha permitido obtener mejores resultados ampliando las bases de datos.

Las técnicas de aprendizaje profundo requieren grandes cantidades de datos, sin embargo las bases de datos con ejemplos de escenarios para vehículos autónomos son escasas y tienen un número reducido de muestras. Por esta razón, hemos llevado a cabo un conjunto de experimentos combinando dos bases de datos. Donde además, hemos probado la efectividad de los modelos de entrenamiento usando etiquetas correctas como referencias parcialmente disponibles, como consecuencia de combinar bases de datos destinadas a diferentes aplicaciones. Los siguientes resultados y un video adjunto muestran una mejora significativa en nuestro caso de ejemplo, por tanto se abre así una nueva forma de mejorar el rendimiento de las detecciones de forma independiente de la arquitectura del detector (Figuras 13 y 14).

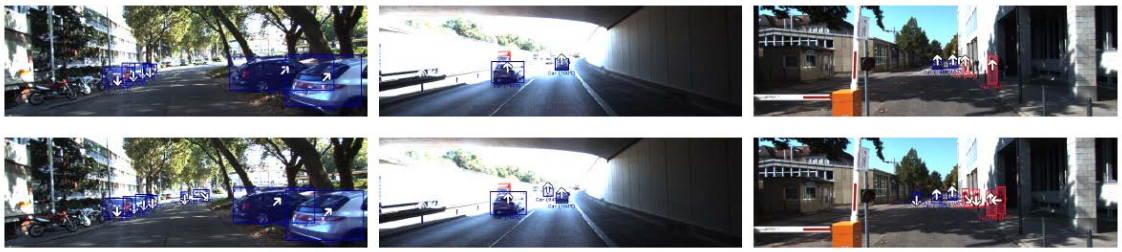


Fig. 13. Ejemplos seleccionados de resultados de detección de objetos y estimación de ángulos de orientación utilizando la base de datos de prueba KITTI. Fila superior: modelo entrenado con

el conjunto de datos de entrenamiento del KITTI; fila inferior: modelo entrenado con el conjunto de datos combinado.

Fig. 14. Ejemplos seleccionados de resultados de detección de objetos y de estimación de ángulos de orientación, utilizando el conjunto de datos de prueba KITTI con categorías adicionales.

Referencias

1. Broggi, A., Cerri, P., Debattisti, S., Laghi, M.C., Medici, P., Panciroli, M., Prioletti, A.: PROUD-public road urban driverless test: architecture and results. In: Proc. IEEE Intelligent Vehicles Symposium (IV). (2014) 648{654

2. Zhu, H., Yuen, K.V., Mihaylova, L., Leung, H.: Overview of Environment Perception for Intelligent Vehicles. IEEE Transactions on Intelligent Transportation Systems (2017)

3.Franke,U., Pfeier, D., Rabe, C., Knoeppel, C., Enzweiler, M., Stein, F., Herrtwich, R.G.: Making Bertha See. In: IEEE International Conference on Computer Vision Workshops (ICCVW). (2013) 214{221

4. Musleh, B., de la Escalera, A., Armingol, J.M.: U-V disparity analysis in urban environments. In: Computer Aided Systems Theory - EUROCAST 2011. Springer Berlin Heidelberg (2012) 426{432

5. Badino, H., Franke, U., Mester, R.: Free space computation using stochastic occupancy grids and dynamic programming. In: IEEE International Conference on Computer Vision Workshops (ICCVW). (2007)

6. Oniga, F., Nedevschi, S.: Processing dense stereo data using elevation maps: Road surface, traffic isle, and obstacle detection. IEEE Transactions on Vehicular Technology 59(3) (2010) 1172{1182

7. Broggi, A., Cattani, S., Patander, M., Sabbatelli, M., Zani, P.: A full-3D Voxel-based Dynamic Obstacle Detection for Urban Scenario using Stereo Vision. In: Proc. IEEE International Conference on Intelligent Transportation Systems (ITSC). (2013) 71{76

8. Felzenszwalb, P.F., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9) (2010) 1627{1645

9. Tian, W., Lauer, M.: Fast Cyclist Detection by Cascaded Detector and Geometric Constraint. In: Proc. IEEE International Conference on Intelligent Transportation Systems (ITSC). (2015) 1286{1291

10. Pepik, B., Stark, M., Gehler, P., Schiele, B.: Teaching 3D geometry to deformable part models. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 3362{3369

11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Proc. Advances in Neural Information Processing Systems (NIPS). (2012) 1097{1105

12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) 580{587

13. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6) (2016) 1137{1149

14. Li, J., Mei, X., Prokhorov, D.: Deep Neural Network for Structural Prediction and Lane Detection in Traffic Scene. IEEE Transactions on Neural Networks and Learning Systems 28(3) (2017) 690{703

15. Yang, F., Choi, W., Lin, Y.: Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classiffers. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
(2016) 2129{2137

16. Yang, L., Liu, J., Tang, X.: Object detection and viewpoint estimation with automasking neural network. In: Computer Vision - ECCV 2014. (2014) 441{455

17. Tulsiani, S., Malik, J.: Viewpoints and Keypoints. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 1510{1519

18. Martn, D., Garca, F., Musleh, B., Olmeda, D., Pelaez, G.A., Marn, P., Ponz, A., Rodrguez Garavito, C.H., Al-Kaff, A., de la Escalera, A., Armingol, J.M.: IVVI 2.0: An intelligent vehicle based on computational perception. Expert Systems with Applications 41(17) (2014) 7927{7944

19. Guindel, C., Martin, D., Armingol, J.M.: Joint object detection and viewpoint estimation using CNN features. In: Proc. IEEE International Conference on Vehicular Electronics and Safety (ICVES). (2017) 145{150

20. Girshick, R.: Fast R-CNN. In: Proc. IEEE International Conference on Computer Vision (ICCV). (2015) 1440{1448

21. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115(3) (2015) 211{252

22. Hirschm□uller, H.: Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2) (2008) 328{341

23. Fischler, M.A., Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the ACM 24(6) (1981) 381{395

24. de la Escalera, A., Izquierdo, E., Martn, D., Musleh, B., Garca, F., Armingol, J.M.: Stereo visual odometry in urban environments based on detecting ground features. Robotics and Autonomous Systems 80(June) (2016) 1{10

25. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 3354{3361

26. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision - ECCV 2014. Springer International Publishing (2014) 818{833

27. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR abs/1409.1 (2014)

28. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama,S., Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding. In: Proc. ACM International Conference on Multimedia. (2014) 675{678