ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA EN ELECTRÓNICA, AUTOMATIZACIÓN Y CONTROL

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN ELECTRÓNICA, AUTOMATIZACIÓN Y CONTROL

TEMA: SISTEMA AUTOMÁTICO DE DETECCIÓN DE PEATONES EN LA NOCHE USANDO INFORMACIÓN VISUAL EN EL INFRARROJO LEJANO BASADO EN REDES NEURONALES CONVOLUCIONALES

AUTOR: BARRENO REYES, LUIS MIGUEL

DIRECTOR: Dr. FLORES CALERO, MARCO JAVIER

SANGOLQUÍ

2017

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA EN ELECTRÓNICA, AUTOMATIZACIÓN Y CONTROL

CERTIFICACIÓN

Certifico que el trabajo de titulación, ***"SISTEMA AUTOMÁTICO DE DETECCIÓN DE PEATONES EN LA NOCHE USANDO INFORMACIÓN VISUAL EN EL INFRARROJO LEJANO BASADO EN REDES NEURONALES CONVOLUCIONALES"*** realizado por el señor ***BARRENO REYES LUIS MIGUEL***, ha sido revisado en su totalidad y analizado por el software anti-plagio, el mismo cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, por lo tanto me permito acreditarlo y autorizar al señor ***BARRENO REYES LUIS MIGUEL*** para que lo sustenten públicamente.

Sangolquí, 1 de agosto de 2017

Dr. Marco Javier Flores Calero
DIRECTOR

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

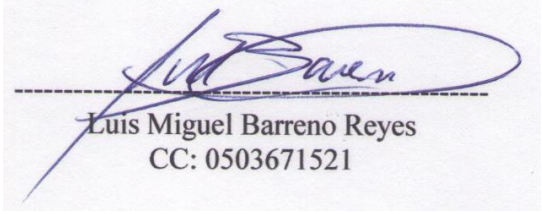
CARRERA DE INGENIERÍA EN ELECTRÓNICA, AUTOMATIZACIÓN Y CONTROL

AUTORÍA DE RESPONSABILIDAD

Yo, ***LUIS MIGUEL BARRENO REYES***, con cédula de identidad N° 0503671521, declaro que este trabajo de titulación ***"SISTEMA AUTOMÁTICO DE DETECCIÓN DE PEATONES EN LA NOCHE USANDO INFORMACIÓN VISUAL EN EL INFRARROJO LEJANO BASADO EN REDES NEURONALES CONVOLUCIONALES"*** ha sido desarrollado considerando los métodos de investigación existentes, así como también se ha respetado los derechos intelectuales de terceros considerándose en las citas bibliográficas.

Consecuentemente declaro que este trabajo es de mi autoría, en virtud de ello me declaro responsable del contenido, veracidad y alcance de la investigación mencionada.

Sangolquí, 1 de agosto de 2017



DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA EN ELECTRÓNICA, AUTOMATIZACIÓN Y CONTROL

AUTORIZACIÓN

Yo, ***BARRENO REYES LUIS MIGUEL***, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar en la biblioteca Virtual de la institución el presente trabajo de titulación ***"SISTEMA AUTOMÁTICO DE DETECCIÓN DE PEATONES EN LA NOCHE USANDO INFORMACIÓN VISUAL EN EL INFRARROJO LEJANO BASADO EN REDES NEURONALES CONVOLUCIONALES"*** cuyo contenido, ideas y criterios son de nuestra autoría y responsabilidad.

Sangolquí, 1 de agosto de 2017

DEDICATORIA

A Luis y Mariana los mejores padres que alguien pudiera tener, por la fe y la confianza que cada día depositan en mí.

A mi tía Blanquita por todo su apoyo para conmigo en el día a día.

A mi hermana por toda su ayuda, consejos y paciencia convirtiéndose en un estímulo importante en la culminación de mi carrera.

Este trabajo está dedicado a ustedes con cariño y gratitud.

AGRADECIMIENTO

A Dios, por la protección que me ha brindado durante el transcurso de toda la carrera universitaria.

A mis padres, que son mi vida, por el apoyo incondicional, por creer en mí siempre y apoyarme en todos los momentos de mi vida.

A todos mis maestros, cuyas enseñanzas directa o indirectamente me ayudaron en la culminación de este proyecto.

De manera especial al Doctor Marco Flores que con su guía constante fue un pilar fundamental en la finalización de este proyecto.

A todas las personas que me ayudaron de una u otra forma, ustedes saben quiénes son, muchas gracias.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Los sistemas de detección de peatones en la noche permiten al conductor estar al tanto del entorno donde se encuentra, especialmente en ambientes donde la visibilidad es muy baja. Pero, estos sistemas aún están lejos de ser perfectos debido a problemas como pobre adquisición de imágenes, amplia variabilidad en la forma de los peatones, entornos al aire libre con alta variabilidad en iluminación, entre otros.

Esta tesis propone el desarrollo e implementación de un sistema de detección de peatones en la noche, utilizando imágenes en el infrarrojo lejano. Este sistema incluye dos etapas, generación de regiones de interés (ROIs, por sus siglas en inglés Región Of Interest) y reconocimiento de peatones a través de una nueva arquitectura de *fast* R-CNN (del inglés Regions with Convolutional Neural Network) que es uno de los enfoques más usados en aprendizaje profundo. La arquitectura *fast* R-CNN consta de dos partes, una es el generador de características y otra es un clasificador basado en regresión logística. Donde se calculan los parámetros de distribución de probabilidad sobre las categorías peatón y no peatón.

Esta nueva arquitectura ha sido evaluada sobre la base de datos LSIFIR y ha demostrado que la tasa de error del 25.5% es competitiva al compararla con estudios anteriores.

Palabras clave

- **ACCIDENTES DE TRÁFICO**
- **PEATONES**
- **FAST R-CNN**
- **INFRARROJO LEJANO**
- **ROI POOL**

ABSTRACT

Pedestrian detection systems at night allow you to be aware of the environment where you are, especially in environments where visibility is very poor. But, these systems are still far from perfect due to problems such as poor image acquisition, wide variability in the shape of pedestrians, outdoor environments with high variability in lighting, among others.

This investigation proposes the development and implementation of a pedestrian detection system at night, using images in the far infrared. This system includes two modules, region of interest (ROI) generation and pedestrian recognition through a new fast R-CNN architecture. The fast R-CNN architecture consists of two parts, one is the generator of features and the other is a logistic regression classifier that is constructed using these characteristics. Logistic regression generates a probability distribution on the pedestrian and non-pedestrian categories.

This new architecture has been evaluated on the LSIFIR database and has demonstrated that miss rate of 25.5% is competitive when compared to previous studies.

Keywords

- **TRAFFIC ACCIDENTS**
- **PEDESTRIAN**
- **FAST R-CNN**
- **FAR INFRARED**
- **ROI POOL**

CAPÍTULO I
1. INTRODUCCIÓN

1.1 Antecedentes

La detección de peatones se ha vuelto un tema de interés en los últimos años, como parte de los sistemas avanzados de asistencia al conductor (ADAS, por sus siglas en inglés Advanced Driver Assitance Systems) los cuales además incluyen otras formas de asistencia al conductor como control de navegación, asistencia de parqueo automatizado, reconocimiento de señales de tránsito, detección de somnolencia, alerta de pre colisión, entre otros. Estos sistemas, aunque algunos son más prescindibles que otros, todos tienen un objetivo común el de ayudar al conductor a mejorar su confort, aumentar la seguridad vial asistiendo y alertando al conductor de posibles situaciones peligrosas que se pueden presentar y en ciertos casos incluso tomar el control parcial del vehículo.

Los sistemas ADAS forman parte de la nueva tendencia de seguridad integrada la cual empezó inicialmente desde un enfoque llamado de seguridad pasiva el cual incluye sistemas que reducen el daño físico en el conductor ya que actúan después de la colisión tal como se observa en la **Figura 1**. La línea vertical anaranjada indica el momento de colisión, a la izquierda se representan sistemas que actúan antes de la colisión y a la derecha después de la colisión. Una de las características principales de los sistemas ADAS es que actúan previo a la colisión, con lo cual ayudan al conductor a identificar potenciales riesgos y a tomar medidas correctivas que son llevadas a cabo en un tiempo entre los 2s y 10s antes de la colisión. Varios estudios prueban que el 70% de accidentes graves podrían haberse evitado mediante el uso de sistemas de asistencia al conductor (Li & Zhu, 2013). Los sistemas ADAS son muy diferentes a cualquier sistema convencional de lazo cerrado que dependa de datos objetivos de algún sensor, los sistemas ADAS dependen de datos de sensores y de datos subjetivos asociados al comportamiento y estado emocional del conductor (Li & Zhu, 2013).

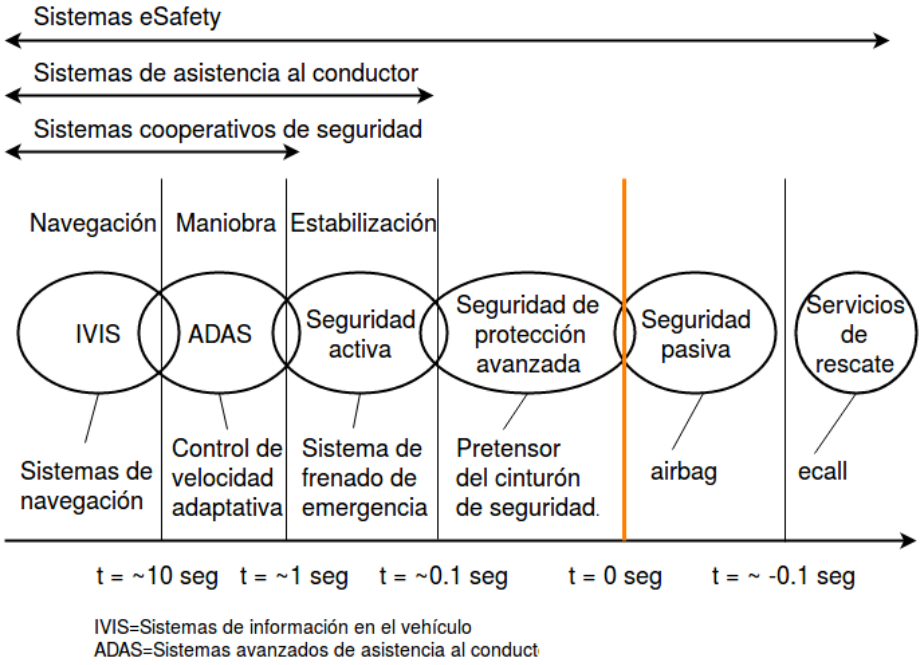


Figura 1. Concepto de Seguridad Integrada.

Fuente: (Li & Zhu, 2013)

Los sistemas de detección de peatones contribuyen a la protección del actor más vulnerable, el peatón, al prevenir accidentes en la vía pública, aunque debido a varios factores como ambientes urbanos con alta variabilidad, condiciones ambientales cambiantes, vibraciones propias del vehículo o por una pobre adquisición de imágenes estamos todavía lejos de obtener sistemas cien por ciento confiables.

En el presente trabajo se desarrollará un sistema de detección de peatones exclusivamente en la noche ya que es donde ocurre la mayor cantidad de accidentes que involucran peatones comparado con los que ocurren durante el día (Soga, Hiratsuka, Fukamachi, & Ninomiya, 2008).

1.1.1 Accidentes de tránsito

Según el diccionario de la Real Academia Española, un accidente se define como un suceso eventual o acción de que resulta daño involuntario para las personas o las cosas, en (Los accidentes de tránsito son un problema socioeconomico, s.f.) específicamente se define a los accidentes de tránsito como un suceso negativo producido por un vehículo en circulación o un peatón, con resultado de daños materiales y/o lesiones o

muertes. Estas dos definiciones tienen en común que ambas explican claramente que se trata de un suceso totalmente involuntario no premeditado, legalmente si no fuera así, se estaría hablando de algo totalmente distinto, un delito penal.

Dependiendo del daño generado durante un accidente de tránsito este puede ser calificado como infracción de tránsito y según la Ley Orgánica de Transporte Terrestre, Tránsito y Seguridad Vial del Ecuador en su artículo número 106 este se define como las acciones u omisiones que, pudiendo y debiendo ser previstas pero no queridas por el causante, se verifican por negligencia, imprudencia, impericia o por inobservancia de las leyes, reglamentos, resoluciones y demás regulaciones de tránsito.

1.1.1.1 Accidentes de tránsito en el mundo

Anualmente se pierden aproximadamente 1.25 millones de vidas en el mundo a consecuencia de los accidentes de tránsito, de este número 20 a 50 millones de personas sufren de traumatismos no mortales. Del total de personas que mueren en una accidente de tránsito la mitad son “usuarios vulnerables de la vía pública”, es decir, peatones, ciclistas y motociclistas (OMS, Lesiones causadas por el tránsito).

En adición a las lesiones y muertes causadas por los accidentes de tránsito también están las pérdidas económicas a consecuencia de los costos del tratamiento, pérdida o disminución de la productividad por parte de quienes resultan muertos o lastimados o por aquellos miembros de la familia que dejan de lado sus actividades cotidianas porque tienen que atender a sus seres queridos lesionados, investigaciones del año 2010 indican que los accidentes de tránsito le cuestan a un país aproximadamente 3% de su producto interno bruto (OMS, Lesiones causadas por el tránsito).

La OMS declaro que si no se aplica ninguna medida para reducir o evitar accidentes de tránsito, se espera que para 2030 estos se conviertan en la séptima causa de muerte en el mundo, es por esto que la Agenda de Desarrollo Sostenible para 2020

se ha fijado como objetivo reducir a la mitad el número de muertos y lesionados por accidentes de tránsito (OMS, Lesiones causadas por el tránsito).

1.1.1.2 Estado actual de la seguridad vial en el mundo

El mundo ha visto una estabilización en el número de muertes por accidentes de tránsito desde el 2013, esto pese al aumento tanto de la población mundial, como del uso de vehículos de motor, específicamente entre 2010 y 2013 la población ha aumentado un 4% y los vehículos un 16%, esta estabilización indica que las campañas de prevención y concientización están funcionando (OMS, Informe sobre la situación mundial de la seguridad vial 2015, 2015). Esta estadística no es del todo alentadora y es que solo 69 países han presentado un descenso en el número de muertes por accidentes de tránsito mientras que 68 países han presentado un aumento del número de defunciones por esta misma causa, lamentablemente 84% de estos países son de ingresos medios y bajos (OMS, Informe sobre la situación mundial de la seguridad vial 2015, 2015).

No se tiene del todo claro por qué los países de ingresos medios y bajos presentan más accidentes de tránsito (aproximadamente el doble) comparada con el número de decesos registrados en los países de ingresos altos. Es la región de África la que encabeza la lista con la mayor tasa de mortalidad por accidentes de tránsito mientras que los países más ricos de Europa están a la cola de esta estadística tal como se muestra en la **Figura 4** (OMS, Informe sobre la situación mundial de la seguridad vial 2015, 2015).

Del total de muertes causadas por accidentes de tránsito aproximadamente la mitad son: peatones (22%), ciclista (4%) y motociclistas (23%) tal como se muestra en la **Figura 5**. La probabilidad de estos usuarios de perder la vida en un accidente de tránsito depende de las acciones preventivas que se tengan en cada región (OMS, Informe sobre la situación mundial de la seguridad vial 2015, 2015).

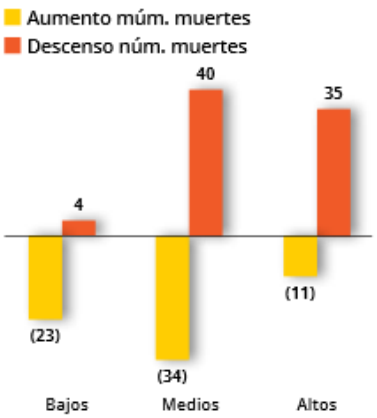


Figura 2. Países que han registrado cambios en el número de muertes por accidentes de tránsito, 2010-2013, por nivel de ingresos.
Fuente: (OMS, Informe sobre la situación mundial de la seguridad vial 2015, 2015)

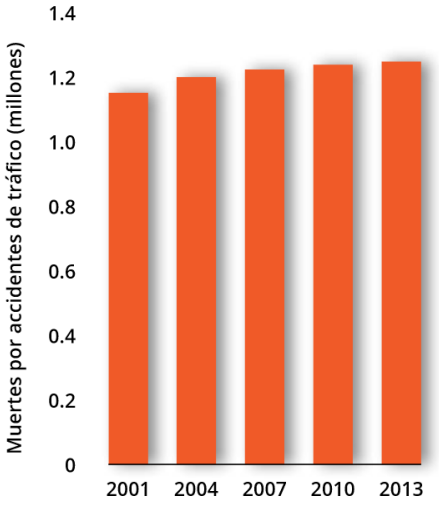


Figura 3. Número de muertes por accidentes de tránsito en el mundo.
Fuente: (OMS, Informe sobre la situación mundial de la seguridad vial 2015, 2015)

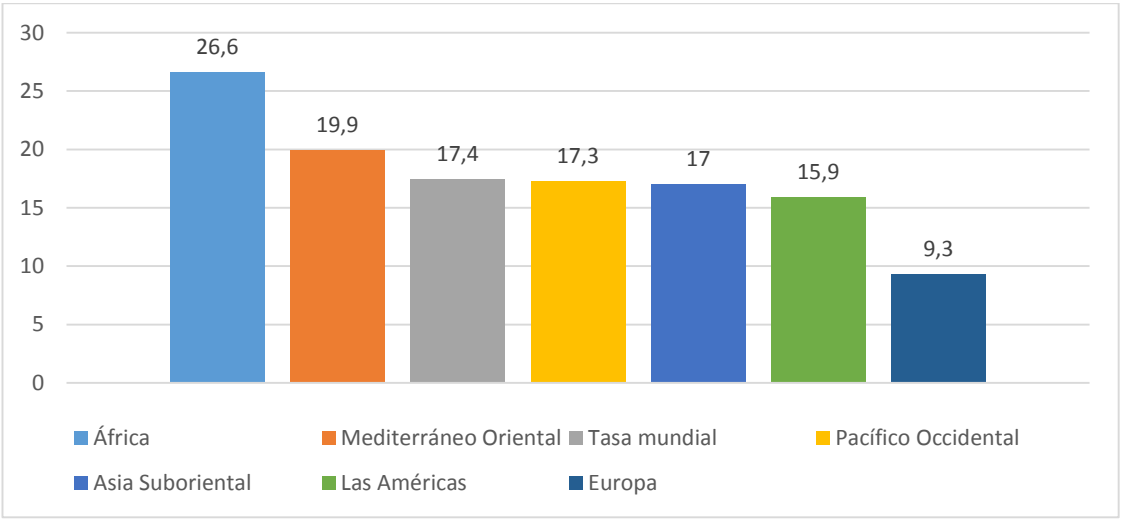


Figura 4. Tasa de mortalidad por accidentes de tránsito por cada 100 000 habitantes.
Fuente: (OMS, Informe sobre la situación mundial de la seguridad vial 2015, 2015)

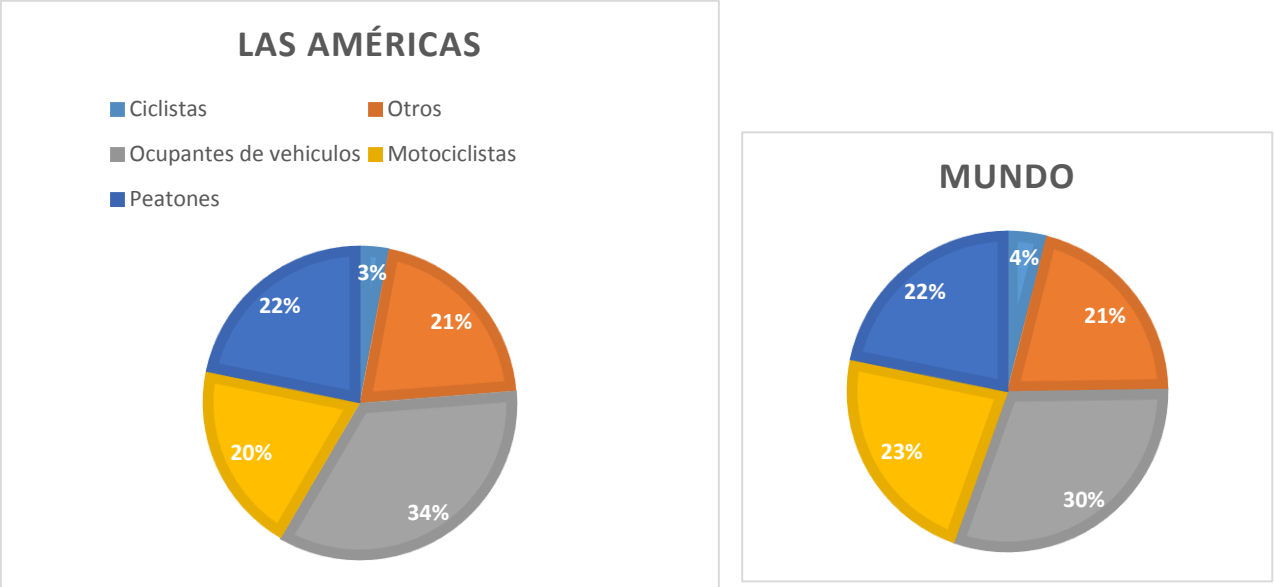


Figura 5. Muertes por accidentes de tránsito en función del tipo de usuario (2013). Fuente: (OMS, Informe sobre la situación mundial de la seguridad vial 2015, 2015)

También es necesario conocer cuáles son los factores de riesgos clave que causan accidentes de tránsito y directamente producen pérdidas humanas y materiales (OMS, Informe sobre la situación mundial de la seguridad vial 2015, 2015):

- Velocidad excesiva.
- Conducción bajo los efectos del alcohol.
- No uso de cascos de motocicleta.
- No uso de cinturones de seguridad y medios de sujeción de niños.
- Distracciones al conducir.

El informe sobre la situación mundial de la seguridad vial de 2015 define muy claramente los factores de riesgos y la forma de afrontarlos:

> Los traumatismos por accidentes de tránsito pueden prevenirse. Los gobiernos tienen que adoptar medidas para abordar la seguridad vial de una forma integral, lo que requiere la participación de muchos sectores entre ellos los propios usuarios de la vía pública.
>
> Entre las intervenciones eficaces cabe mencionar el diseño de una infraestructura más segura y la incorporación de elementos de seguridad al decidir el uso de la

tierra y planificar el transporte; el **mejoramiento de los elementos de seguridad de los vehículos**; y la atención mejorada de las víctimas inmediatamente después de los accidentes de tránsito (OMS, Informe sobre la situación mundial de la seguridad vial 2015, 2015).

1.1.1.3 Accidentes de tránsito en el Ecuador

Según datos del INEC en el 2014 los accidentes de transporte terrestre representan la sexta causa de muerte en el país tal como se muestra en la **Tabla 1**, debido probablemente a la poca concientización que se tiene, situación que a nivel gubernamental ha causado gran preocupación por lo que el gobierno se ha visto en la necesidad de realizar grandes campañas de prevención y concienciación, especialmente en los feriados que es cuando la población se moviliza mucho más.

La Agencia Nación de Transito (ANT) en su publicación del mes de diciembre de 2015 informa que se tuvieron en ese año 35706 accidentes, en contraste con los 38658 obtenidos en el año 2014 tal como se observa en la **Figura 6**. La última publicación que se tiene disponible, la de octubre de 2016 muestra que hasta ese mes se tienen 25044 accidentes. Como resultado de estos accidentes se tienen muertos y lesionados, respecto a las muertes en 2015 se tuvieron 2138 muertes a diferencia de los 2322 que se tuvieron en 2014, mientras que hasta octubre de 2016 se tienen 1607 muertes, respecto a lesionados, en 2014 se tuvieron 27668 lesionados mientras que en 2015 se tuvo un ligero descenso, 25234 lesionados finalmente hasta octubre de 2016 se tienen 17532 lesionados. En conclusión se puede ver una ligera tendencia a la baja de accidentes desde el 2014 y en consecuencia de muertos y lesionados (ANT, Estadisticas de Transporte Terrestre y Seguridad Vial, 2016).

Tabla 1.
Principales causas de mortalidad general de 2014.

Causas de muerte	**Numero**	**%**
Enfermedades isquémicas del corazón	4430	7.03

Continúa

Diabetes Mellitus	4401	6.99
Enfermedades cerebrovasculares	3777	6.00
Enfermedades hipertensivas	3572	5.67
Influenza y neumonía	3418	5.43
Accidentes de transporte terrestre	3059	4.86
Población estimada 2014	16 027 466	
Total de defunciones	62 981	
Tasa de mortalidad general (x 100000 hab.)	39 296	

Fuente: (INEC, 2015)

En 2015 la causa más probable de accidentes en el Ecuador con 13.7% fue no respetar las señales reglamentarias de tránsito, mientras que hasta octubre 2016 la causa más probable con 21% es conducir desatento a las condiciones de tránsito (celular, pantallas de video, comida, maquillaje o cualquier otro elemento distractor). La tercera causa de muerte más probable es por no ceder el derecho de vía o preferencia de paso al peatón con 7.6%.

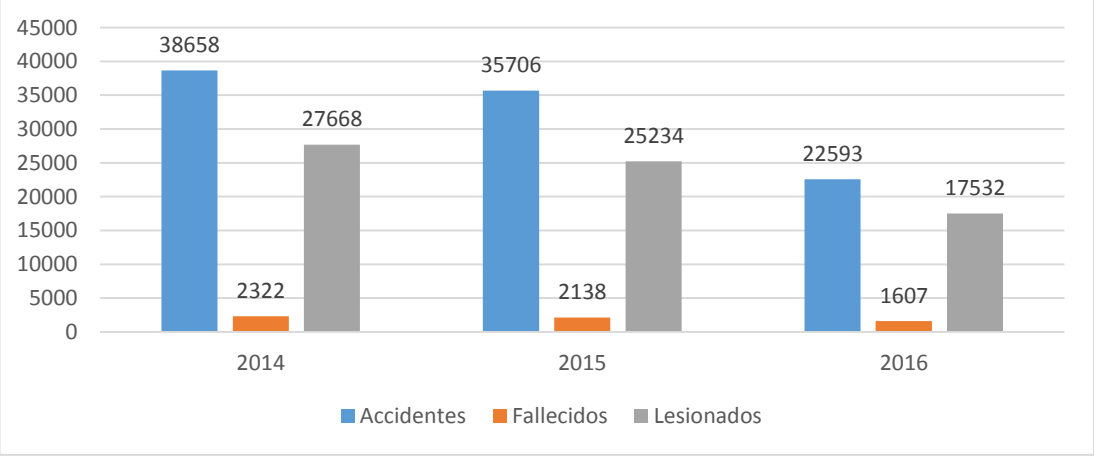


Figura 6. Número de accidentes, fallecidos y lesionados desde 2014.

Fuente: (ANT, Estadisticas de Transporte Terrestre y Seguridad Vial, 2016).

El tipo de accidente de tránsito que más ocurre en el país, como se puede ver en la **Tabla 2** es el choque lateral con 6936 seguido por el atropello con 3907, esto hasta octubre de 2016.

Tabla 2.
Siniestros de transito por tipo a nivel nacional Octubre 2016.

TIPO	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	TOTAL	%
CHOQUE LATERAL	888	739	675	676	708	633	723	648	572	674	**6.936**	**27,70**
ATROPELLO	430	402	418	396	358	386	376	383	342	416	**3.907**	**15,60**
ESTRELLAMIENTO	429	351	326	313	320	299	338	325	293	315	**3.309**	**13,21**
PERDIDA DE PISTA	339	304	257	262	221	250	273	267	244	263	**2.680**	**10,70**
CHOQUE POSTERIOR	320	278	265	276	269	231	273	237	202	238	**2.589**	**10,34**
ROZAMIENTO	168	186	141	171	154	169	184	159	165	182	**1.679**	**6,70**
CHOQUE FRONTAL	155	135	133	127	125	121	111	97	104	110	**1.218**	**4,86**
COLISION	87	88	81	57	71	53	72	70	51	64	**694**	**2,77**
CAIDA DE PASAJERO	56	61	78	59	70	74	75	78	71	53	**675**	**2,70**
VOLCAMIENTO	106	66	65	68	61	58	62	65	65	56	**672**	**2,68**
OTROS	43	47	53	34	38	42	47	44	37	45	**430**	**1,72**
ARROLLAMIENTO	23	22	21	14	30	43	25	23	19	35	**255**	**1,02**
TOTAL	3.044	2.679	2.513	2.453	2.425	2.359	2.559	2.396	2.165	2.451	**25.044**	**100**

Fuente: (ANT, Siniestros octubre 2016, 2016)

1.1.2 Sistemas inteligentes de transporte

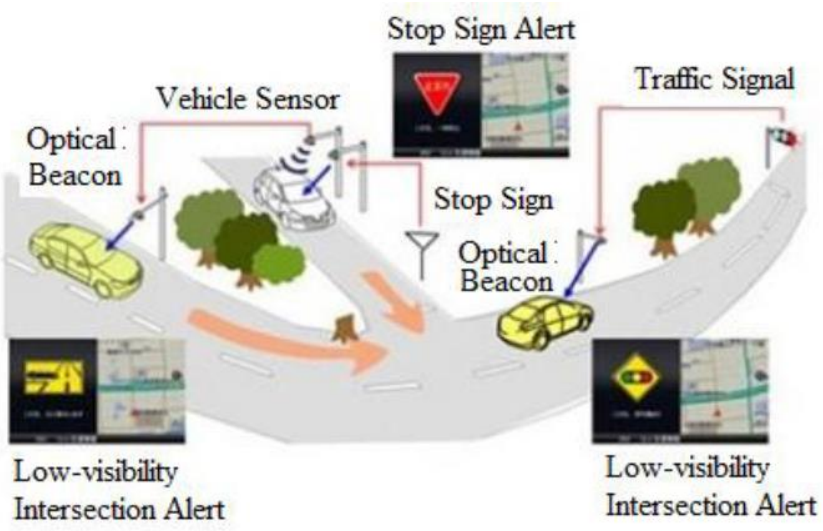
Los sistemas inteligentes de transporte (ITS, por sus siglas en inglés Intelligent Transportation System) “aplican tecnologías avanzadas de electrónica, comunicación, computación, control y sensores y detección en toda clase de sistemas de transportación, esto con el fin de mejorar la seguridad, eficiencia y servicio, y trafico mediante la trasmisión de información en tiempo real” (Brief introduction to Intelligent Transportation System, ITS, s.f.). Los ITSs tienen como objetivos:

- Mejorar la seguridad vehicular.
- Aliviar la congestión vehicular.
- Mejorar la eficiencia en la trasportación.
- Reducir la contaminaron atmosférica.
- Incrementar la eficiencia energética.
- Promover el desarrollo de la industria automotriz y de las industrias relacionadas.

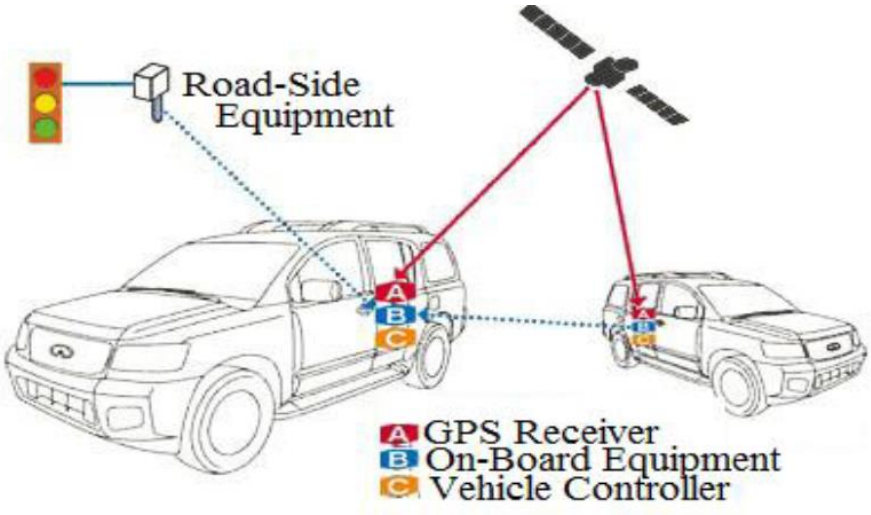
La clasificación más usada que se tiene en ITS está basada en el posicionamiento de los sistemas (Vanajakshi, Ramadurai, & Anand, 2010) y son las siguientes:

- **A nivel vehicular:** Tecnologías desplegadas dentro del vehículo, como sensores, displays, procesadores, entre otros, que sirven para proveer información al conductor.
- **A nivel de infraestructura:** Tecnologías que están desplegadas en la carretera y que recogen información sobre el tráfico.
- **A nivel cooperativo:** Comunicación tanto a nivel vehicular como también entre los niveles vehicular y de infraestructura.

a)



b)



c)

Figura 7. Clasificación ITS basada en el posicionamiento de los sistemas. a) A nivel vehicular, b) A nivel de infraestructura, c) A nivel cooperativo. Fuente: (Vanajakshi, Ramadurai, & Anand, 2010)

Aunque según su funcionalidad los ITSs también se clasifican en 6 categorías (Brief introduction to Intelligent Transportation System, ITS, s.f.):

- **Sistemas avanzados de gestión de tráfico:** Integra varios subsistemas que responden en tiempo real a las condiciones cambiantes del tráfico mediante una sola interface que provee información sobre este, además predice el estado del tráfico para una eficiente planificación.
- **Sistemas avanzados de información al viajero:** Permite al viajero acceder a información que lo asisten en el proceso de toma de decisión respecto a qué modo de transporte elegir, tiempo de viaje, elección de rutas, evasión de congestionamiento.
- **Sistemas avanzados de control vehicular y seguridad:** Ayuda al conductor a tener un mejor control vehicular para evitar accidentes, incluye sistemas de alerta de pre colisión, sistemas automáticos de frenado o aceleramiento, sistemas de parqueo automático, entre otros.
- **Sistemas avanzados de transporte público:** Utiliza las tres categorías mencionadas anteriormente con el fin de mejorar la calidad del servicio de transporte e incentivar el uso de transporte público.
- **Operaciones de vehículos comerciales:** Al igual que el anterior utiliza las tres primeras clasificaciones inicialmente mencionadas en operación de vehículos comerciales con el objetivo de mejorar la eficiencia y seguridad.
- **Sistemas avanzados de asistencia a la conducción:** Integración de varias tecnologías que apoyan al conductor en el difícil proceso de la conducción.

1.1.2.1 Sistemas avanzados de asistencia a la conducción

Hammond en (Hammond, Qu, & Rawashdeh, 2015) define a los sistemas avanzados de asistencia a la conducción como “tecnologías que apoyan al conductor en el complejo proceso de controlar un vehículo sin peligro ya sea brindando información al conductor sobre el entorno o tomando el control parcial del vehículo para prevenir accidentes”

Los sistemas ADAS dependen de sensores y cámaras que censan los exteriores del automotor así como el estado del conductor, estos elementos se comunican entre sí mediante la infinidad opciones de conectividad que se tienen ahora (Hammond, Qu, & Rawashdeh, 2015). Algunas de las aplicaciones más conocidas de sistemas ADAS basadas en visión artificial son las siguientes (Hammond, Qu, & Rawashdeh, 2015):

- Sistemas de advertencia de salida de carril.
- Sistemas de asistencia de mantenimiento en carril.
- Sistemas de detección de peatones.
- Reconocimiento de señales de tránsito.
- Advertencia de colisión delantera.
- Sistemas avanzados de iluminación frontal.
- Sistemas de visión circundante.
- Sistemas de visión en la noche.

1.2 Planteamiento del problema

En años recientes la detección de peatones en tiempo real ha capturado mucho la atención no solo en la academia sino también en la industria debido a la gran cantidad de aplicativos que tienen estos sistemas que van desde video vigilancia, vehículos de conducción autónoma, seguridad vehicular, sistemas de recuperación de imágenes basados en contenido, entre otros.

Tradicionalmente existen dos enfoques para afrontar la problemática de detección de peatones, el primero es la detección de peatones basado en visión artificial y el segundo es la detección de peatones basado en medidas activas y pasivas (sensores). Comúnmente este problema se afronta usando visión artificial exclusivamente pero también se pueden combinar con el uso de sensores como ultrasónico, radar o LIDAR para obtener mejores resultados (Hammond, Qu, & Rawashdeh, 2015).

En la detección de peatones basada en visión artificial, uno de los aplicativos ADAS basado en visión artificial más populares son los sistemas de detección de peatones y se define como la aplicación de información visual en la detección y localización de peatones para luego advertir al conductor del peligro, los peatones pueden ser detectados tanto en imágenes del espectro visible como del infrarrojo cercano (NIR, por sus siglas en inglés Near Infrared) o lejano (FIR, por sus siglas en inglés Far Infrared). En los sistemas de detección de peatones NIR se ilumina activamente una escena en el espectro NIR y se captura la radiación reflejada mientras que los sistemas de detección en FIR detectan pasivamente la radiación térmica de los objetos en la escena (Lim, Tsimhoni, & Liu, 2010). Sin embargo, los sistemas de detección de peatones NIR tienen una distancia de detección más corta que la de los sistemas FIR. Por lo tanto, la mayoría de los sistemas de detección de peatones diseñados para uso nocturno emplean imágenes FIR que también se denominan imágenes térmicas (Lin, y otros, 2015).

La imágenes en el espectro visible presentan la desventaja que son muy dependientes de las condiciones de iluminación del ambiente, por lo cual no son adecuadas para la detección de peatones en la noche que es donde ocurre la mayor cantidad de accidentes que involucra peatones comparada con la cantidad de accidentes que ocurre durante el día (Soga, Hiratsuka, Fukamachi, & Ninomiya, 2008). Las imágenes del infrarrojo presentan características deseables para la detección de peatones en la noche ya que no requieren ningún tipo de iluminación y además los peatones resaltan en la imagen debido a que emiten más calor que los objetos que se encuentran comúnmente en un ambiente urbano (Fang, Yamada, Ninomiya, Horn, & Masaki, Comparison between infrared-image-based and visible-image-based approaches for pedestrian detection, 2003).

La seguridad del peatón es un problema que nos afecta a todos, todos caminamos para cumplir con nuestras responsabilidades diarias o algunas veces caminos para ejercitarnos, caminar es el medio de transporte más básico y universal, aunque desgraciadamente debido al aumento poblacional y al aumento de la frecuencia del uso vehicular cada vez es más peligroso ser un peatón en el mundo especialmente en sociedades donde se tiene una inercia de años de irrespeto a las leyes de tránsito. Se

tiene cada vez más vehículos de motor circulando sin embargo las infraestructuras públicas donde se movilizan los peatones siguen siendo las mismas (OMS, Seguridad peatonal, 2013).

En Ecuador la Policía Nacional y la Agencia Nacional de Tránsito (ANT) en conjunto con el Ministerio del Interior trabajan anualmente en campañas de prevención de accidentes de tránsito. Gracias a estas medidas gubernamentales se registró un ligero decremento del número de fallecidos y lesionados en los últimos años pero al tratarse de vidas humanas, este decremento no es suficiente principalmente porque las colisiones con peatones son previsibles y evitables (OMS, Seguridad peatonal, 2013).

Un transporte seguro es vital para el decremento de decesos y lesionados causados por accidentes de tránsito, un transporte seguro que dé cabida al error humano y que tenga en cuenta a los actores más vulnerables de la vía pública como peatones y ciclistas. Ello requiere una política que se centre en las infraestructuras viales, en el diseño vehicular y en la gestión de la velocidad, y que se apoye en una serie de medidas educativas, legislativas, reglamentarias y sancionadoras (OMS, Seguridad peatonal, 2013).

1.3 Justificación e importancia

Para esta investigación es importante conocer durante qué horas del día los accidentes ocurren con mayor frecuencia ya que se parte de la premisa que durante la noche ocurre la mayor cantidad de accidentes, se puede ver en la **Figura 8** que entre las 19:00 y 19:59 y los días sábado ocurre la mayor cantidad de accidentes.

En Ecuador los vehículos en circulación aumentan considerablemente cada año, según (Anuario de Estadísticas de Transportes 2013, 2014) hasta el año 2013 se tenía 1.717.886 vehículos matriculados, la última estadística presentada muestra que hasta el año 2014 se obtuvieron 1.752.712 vehículos matriculados (Anuario de Estadísticas y Transportes 2014, 2015), esta razón en conjunto con las principales causas de accidentes en el país como impericia del conductor, irrespeto a las señales de tránsito

y exceso de velocidad contribuyen a la generación de accidentes que involucran peatones, el número de accidentes por atropellamientos en el año 2014 fue de 5.983 según (Anuario de Estadísticas y Transportes 2014, 2015).

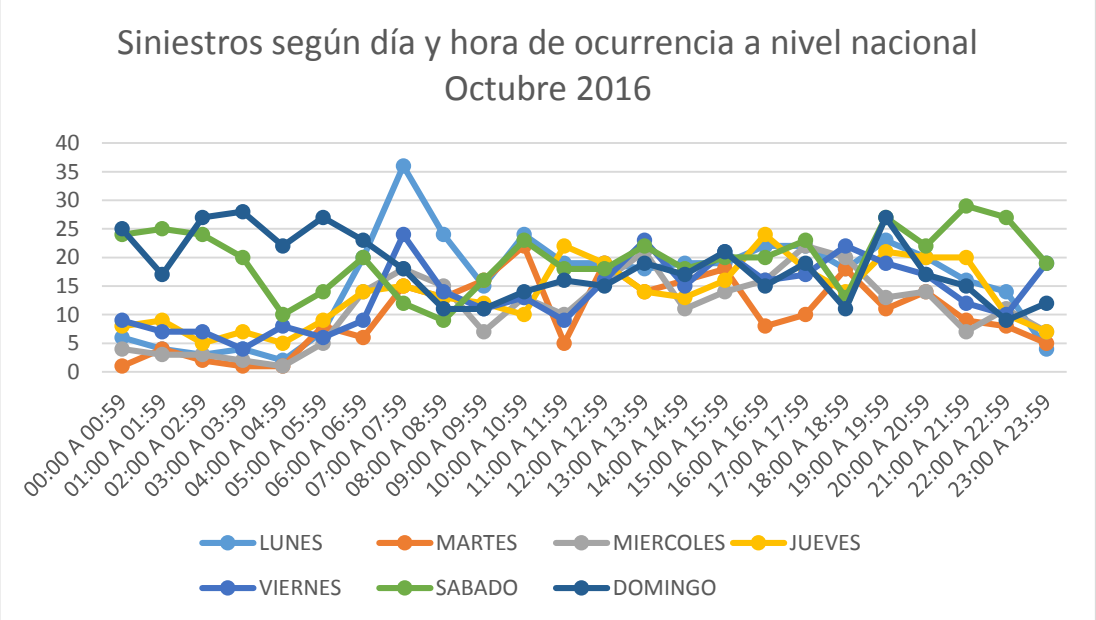


Figura 8. Siniestros según día y hora de ocurrencia a nivel nacional Octubre 2016**.**
Fuente: (ANT, Siniestros octubre 2016, 2016)

A nivel internacional existen nuevas regulaciones de la Unión Europea que exigen que los vehículos incluyan sistemas de protección al peatón, automatizando la luz del vehículo, el sistema de frenos, el sistema de navegación, entre otros (O'Malley, Jones, & Glavin, 2010). A nivel mundial más de 10 millones de personas se encuentran afectadas por accidentes de tránsito y de 2 a 3 millones sufrieron lesiones graves, yendo más allá el número de accidentes en la noche es 4 veces más que durante el día (Qi, John, Liu, & Mita, 2016), añadiendo a esto, los conductores de casi 50 años necesitan más luz para conducir en la noche comparados con los de 30 años por lo que un sistema de visión en el infrarrojo ayudaría mucho a estos conductores ya que permitiría visualizar objetos más allá de lo que cubren las luces delanteras del vehículo (Fang, Yamada, Ninomiya, Horn, & Masaki, Comparison between infrared-image-based and visible-image-based approaches for pedestrian detection, 2003), es por esto que es imperativo que existan sistemas de detección de peatones en la noche.

1.4 Alcance del proyecto

El proyecto contempla el desarrollo de un sistema de detección de peatones para aplicaciones en vehículos inteligentes que incluye un algoritmo de generación de ROIs a través de un método de generación de candidatos sobre imágenes infrarrojas, un método de clasificación automática mediante *fast* R-CNN y la fusión de los dos algoritmos para generar un detector de peatones en la noche. Este sistema está orientado a funcionar en un alcance de 5 m. a 25 m. en la zona de alto riesgo de impacto a peatones como se observa en la **Figura 9**.

Para analizar el desempeño del sistema se realizaran pruebas de funcionamiento y en caso de encontrar anomalías se realizaran las correcciones del caso.

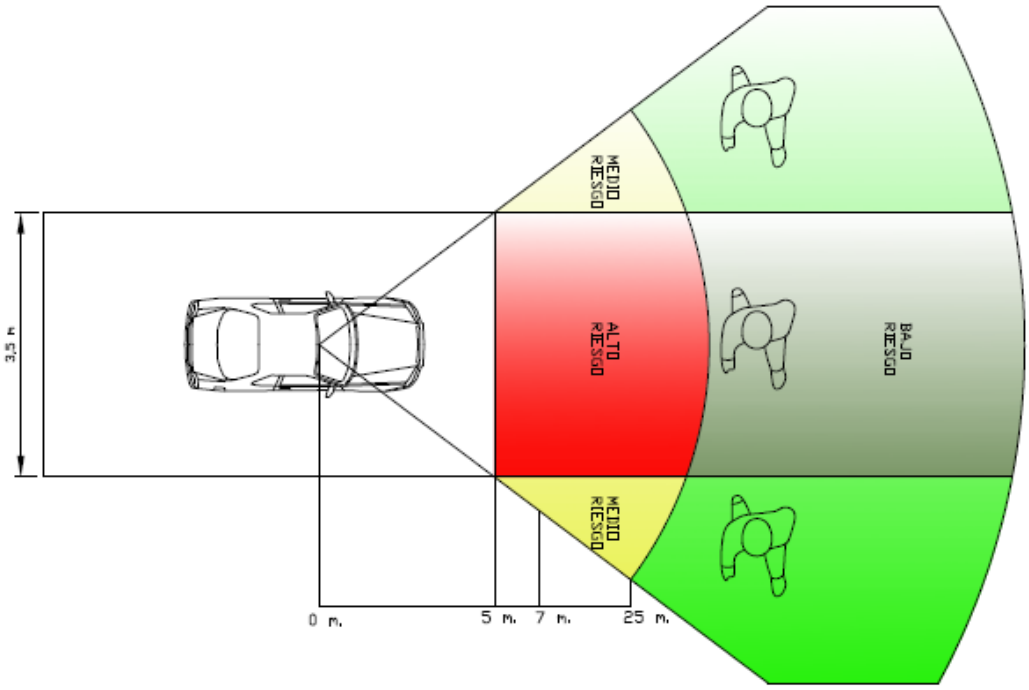


Figura 9. Zonas de riesgo de impacto al peatón.
Fuente: (Delgado Landázuri, 2015)

1.5 Objetivos

1.5.1 Objetivo general

- Desarrollar un sistema automático de detección de peatones en la noche utilizando información visual en el infrarrojo lejano basado en redes neuronales convolucionales, en tiempo real, para sistemas de asistencia a la conducción.

1.5.2 Objetivo específicos

- Desarrollar el estado del arte de las técnicas de detección de peatones en el infrarrojo lejano.
- Desarrollar un algoritmo de generación de candidatos a peatones sobre imágenes tomadas en el infrarrojo lejano, para aplicaciones de tiempo real.
- Desarrollar un algoritmo de clasificación para discriminar entre peatones y no peatones, sobre imágenes en el infrarrojo lejano, mediante redes neuronales convolucionales.
- Realizar pruebas de funcionamiento, en tiempo real, para comprobar el estado de la implementación.

1.6 Resumen de contenidos

El presente trabajo está dividido en seis capítulos. El primer capítulo está dedicado principalmente a la investigación del estado actual de la seguridad vial en el Ecuador y el mundo. El segundo capítulo presenta una revisión del estado del arte y de la técnica en detección de peatones. En el tercer capítulo se define el método de generación de candidatos a utilizarse. El cuarto capítulo trata principalmente de todo lo respecto a la arquitectura *fast* R-CNN a utilizarse. El quinto capítulo presenta resultados individualmente y en conjunto detección y clasificación. En el sexto capítulo se presentan las conclusiones que se encontraron.

CAPÍTULO II
2. ESTADO DEL ARTE

2.1 Introducción

Durante el trascurso de los años se han desarrollado importantes trabajos que han contribuido con el desarrollo que ha tenido la detección de objetos especialmente la detección de peatones. Es importante conocer en qué estado se encuentra la detección de peatones, hasta donde se ha llegado y cuáles son los retos más difíciles a los que se enfrenta esta problemática.

2.2 Bases de datos

2.2.1 Base de datos LSI far infrared pedestrian

Tal como se describe en (Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013) esta base de datos consiste de imágenes FIR recolectadas en ambientes urbanos desde una cámara montada externamente en un vehículo. Estas imágenes fueron adquiridas usando una camera indigo omega bajo una resolución de 164x129 en escala de grises de 14 bits y una longitud focal de 318 pixeles. Una vez recolectadas las imágenes fueron manualmente etiquetadas, indicando las coordenadas que definen el rectángulo delimitador. La base de datos está divida en dos partes (carpetas), una parte para clasificación la cual no es usada en este trabajo y otra para detección.

La parte de detección consta de las imágenes recolectadas directamente con la cámara, las cuales fueron etiquetadas como positivas o negativas, positivas en el caso de que contengan peatones y negativas si no los tiene. Este conjunto consta de 13 escenas, una escena es una secuencia especifica de imágenes seguidas una delante de otra que unidas de cierta forma podrían formar un video, cada una de estas escenas contiene un número variado de imágenes de un solo canal de 14 bits con una dimensión de 164x129 pixeles. En la **Tabla 3** se resume el contenido de esta subcarpeta, donde

el número entre paréntesis representa el número de fotogramas que contiene peatones y el número que esta junto a este representa el número de fotogramas que no contiene peatones.

Tabla 3.
Contenido de la base de datos LSI FIR.

Conjunto de datos	**FIR**	
	Clasificación	Detección
Entrenamiento	43391 (10209)	2936 (3225)
Prueba	22051 (5945)	5788 (3279)

Fuente: (Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013)

2.2.2 Base de datos CVC-09: FIR Sequence Pedestrian

Esta base de datos fue desarrollada por (Socarras, Ramos, Vazquez, Lopez, & Gevers, 2011) y está compuesta de dos conjuntos de imágenes FIR, denominadas DayTime y NightTime según el momento del día el cual fueron adquiridas. El primer conjunto contiene 5990 imágenes y el segundo 5081, la **Tabla 4** resume el contenido de esta base de datos. Los números entre paréntesis especifican el número de ejemplos positivos que tiene cada conjunto y el número junto a este especifica el número de fotogramas de cada conjunto.

Tabla 4.
Contenido de la base de datos CVC-09.

Datasets	**FIR**	
	DayTime	NightTime
Train	3110 (4548)	2198 (4333)
Test	2880 (2304)	2883 (2883)

Fuente: (Socarras, Ramos, Vazquez, Lopez, & Gevers, 2011)

a) b)

Figura 10. Imágenes de las bases de datos a utilizar.
a) Imagen de LSI FIR, b) Imagen de CVC-09: FIR Sequence Pedestrian
Fuente: (Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013) y (Socarras, Ramos, Vazquez, Lopez, & Gevers, 2011).

2.3 Estado de la técnica

El estado de la técnica se divide en dos partes: una destinada a los métodos de generación de regiones de interés (ROIs) y otro destinado a los métodos de clasificación.

2.3.1 Generación de candidatos a peatones

Para la generación de ROIs a peatones en imágenes infrarrojas uno de los primeros enfoques que se utilizó fue el de la ventana deslizante como en (Zhang, Wu, & Nevatia, 2007), este enfoque consiste en seleccionar regiones del fotograma que se generan al hacer pasar una ventana deslizante por el mismo aplicando un desplazamiento en ambos ejes tal como se observa en la **Figura 11**. Este trabajo se repite variando la escala del fotograma o de la ventana. Una de las ventajas de este enfoque es que debido a que se realiza una búsqueda exhaustiva de candidatos es muy difícil que se pase por alto algún ROI, mientras que como desventaja, el recorrer exhaustivamente todo un fotograma y repetir este trabajo a diferentes escalas implica un costo computacional alto y no es adecuado para aplicaciones en tiempo real (Liu, Zhuang, & Ma, 2013).

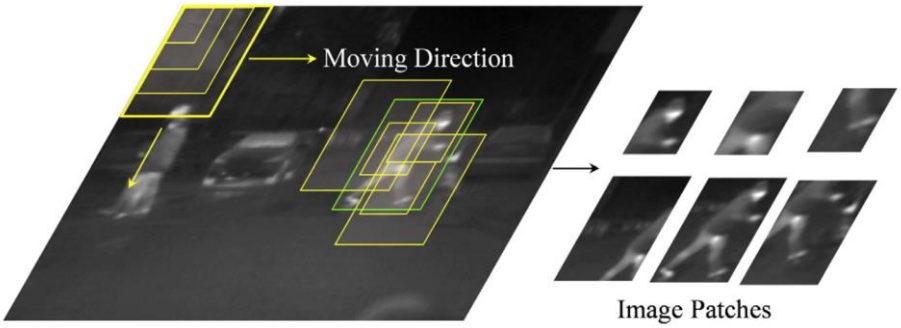


Figura 11. Enfoque ventana deslizante para generación de ROIs en imágenes FIR. Fuente: (Qi, John, Liu, & Mita, 2016)

A partir de aquí utilizando ventana deslizante se ha buscado reducir el número de candidatos generados limitando las regiones de búsqueda en el fotograma como en (Zou, Sun, & Ji, 2012), donde se aplica ventana deslizante solamente en la vecindad de puntos de interés detectados con el algoritmo conocido como SUSAN. Este algoritmo detecta puntos de interés que presentan altas intensidades.

Otro de los métodos que se utilizó para generación de ROIs a peatones en imágenes infrarrojas es el basado en detección de movimiento como en (Lin, y otros, 2015) para lo cual se utilizó una cámara en posición estática donde se restan dos fotogramas sucesivos para detectar movimiento, luego se crea una imagen binaria aplicando un umbral a partir de aquí se selecciona todas las partes que queden con algún método de etiquetado de componentes conectados. Este método tiene la limitante que solo se aplica para cámaras estáticas, es decir este método no es aplicable cuando la cámara esta en movimiento como por ejemplo al estar empotrada en un vehículo otra de las desventajas es que si los peatones no se mueven para nada, al tener dos fotogramas sucesivos totalmente iguales no existiría detección.

Otros enfoques aprovechan la propiedad de que los peatones aparecen más brillantez en las imágenes infrarrojas así que extraen regiones de interés basados en la selección de umbrales de intensidad, este ha sido unos de los métodos de generación de regiones de interés predominantes durante muchos años (Kim & Lee, 2013). En (Suard & Rakotomamonjy, 2006) se escoge manualmente un valor de umbral fijo que permite aislar los objetos más radiantes de la imagen, en (Xu, Liu, & Fujimura, 2005)

se ajusta el umbral dinámicamente según los valores de intensidad medios y máximos de cada fotograma. Los peatones no siempre presentan intensidades uniformes debido a la condiciones físicas de la ropa, además generalmente en los fotogramas se presentan objetos que tienen similar intensidad con los peatones como señales de tránsito, luces peatonales, semáforos, algunas partes del vehículo como el capó y el tubo de escape (debido a su temperatura), entre otros por lo tanto enfoques basados en el umbral de intensidad pueden capturar objetos que no son peatones o incluso pueden solo capturar partes de este aisladamente como son la cabeza, las manos o las piernas, ya que el torso generalmente al estar abrigado presenta menos intensidad que las extremidades o la cabeza. En (O'Malley, Jones, & Glavin, 2010) se presenta un método de compensación debido a las pérdidas de intensidad que se puede tener en el torso causado por el uso de vestimenta de invierno, este método empareja la intensidad de todo el cuerpo, después se genera la región de interés aplicando un valor de umbral alto cercano a la intensidad máxima del fotograma generando puntos semilla (seed points) a partir de aquí se disminuye el umbral periódicamente para que los valores con similares intensidades se vallan uniendo esto mientras la región que va creciendo cumpla con la relación de aspecto y la extensión. Aplicar un solo umbral global a toda una imagen puede dividir a los candidatos en partes aisladas sin conectar, para solucionar esto (Dong, Ge, & Luo, 2007) utiliza dos valores de umbral para cada pixel cuyos valores dependen de la vecindad del pixel. Este método se conoce como segmentación de umbral dual y sus umbrales se calculan con las siguientes ecuaciones:

$$T_{Low}(i,j) = \frac{\sum_{x=i-N}^{i+N} I(x,j)}{L} \tag{1}$$

$$T_{High}(i,j) = T_{Low}(i,j) + \theta \tag{2}$$

Donde $L = 2N + 1$, y L representa el ancho de la vecindad. Una vez que se tienen todos los umbrales de cada pixel, la imagen se segmenta según el siguiente algoritmo:

$$\begin{gathered} p(i,j) \in P, If\ \ I(i,j) > T_{High}(i,j)\ or \\ If\ \ T_{Low} \le I(i,j) \le T_{High}(i,j)\ \ and\ \ p(i-1,j) \in P, \\ p(i,j) \in F, If\ \ I(i,j) < T_{Low}(i,j)\ or \\ If\ \ T_{Low} \le I(i,j) \le T_{High}(i,j)\ \ and\ \ p(i-1,j) \in F \end{gathered} \tag{3}$$

Donde *P* indica peatón y *F* fondo, este algoritmo genera una imagen binaria, resultados de esta segmentación se muestra en **Figura 12**.

a)

b)

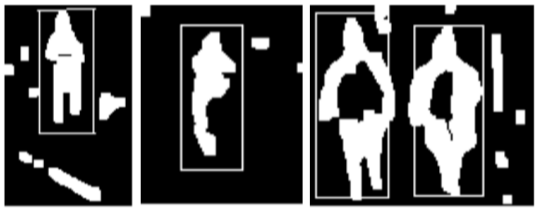


Figura 12. Resultados de segmentación aplicando umbral dual adaptativo: a) imágenes infrarrojas de entrada, b) resultado de segmentación.

Fuente: (Dong, Ge, & Luo, 2007)

a) b)

Figura 13. Falsas detecciones que se dan en ciertos casos al aplicar segmentación de umbral dual adaptivo: a) imágenes infrarrojas de entrada, b) segmentaciones incorrectas.

Fuente: (Ge, Luo, & Tei, 2009)

(Ge, Luo, & Tei, 2009) manifiesta que cuando el fondo y la figura del peatón tienen similares intensidades o cuando dos peatones se encuentran ubicados en proximidad este método tiende a generar falsas detecciones como en **Figura 13**, así que optimiza de cierta forma el cálculo del umbral alto T_{High} con la siguientes ecuaciones:

$$T'_{High}(i,j) = max\{T_1(i,j), T_{Low}(i,j)\} \quad (4)$$

$$T_1(i,j) = min\{T_2(i,j), 230\} \quad (5)$$

$$T_2(i,j) = min\{T_3(i,j), T_{Low}(i,j) + 8\} \quad (6)$$

$$T_3(i,j) = max\{1.6 * (T_{Low}(i,j) - \alpha), T_{Low}(i,j) + 2\} \quad (7)$$

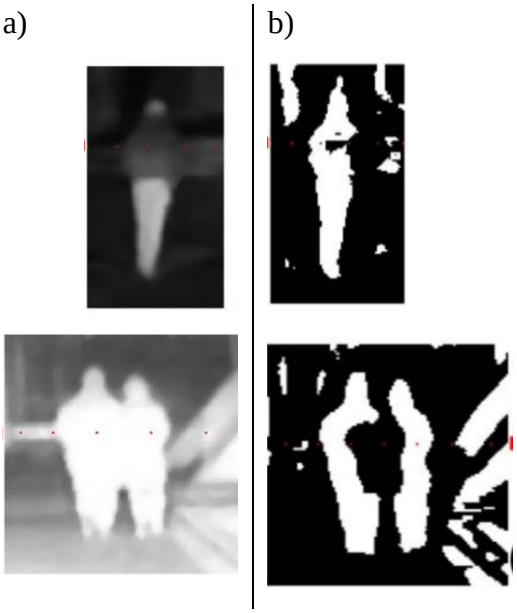


Figura 14. Segmentación obtenida al mejorar el método de umbral dual adaptativo: a) imágenes infrarrojas de entrada, b) candidatos segmentados correctamente.
Fuente: (Ge, Luo, & Tei, 2009)

En (Liu, Zhuang, & Ma, 2013) se generan las regiones de interés mediante la curva de proyección vertical del gradiente, esta curva permite segmentar el fotograma

en franjas verticales indicando las regiones con intensidades más altas luego se aplica segmentación de umbral dual para escoger los candidatos; resultados de candidatos generados con este método se muestran en **Figura 15**.

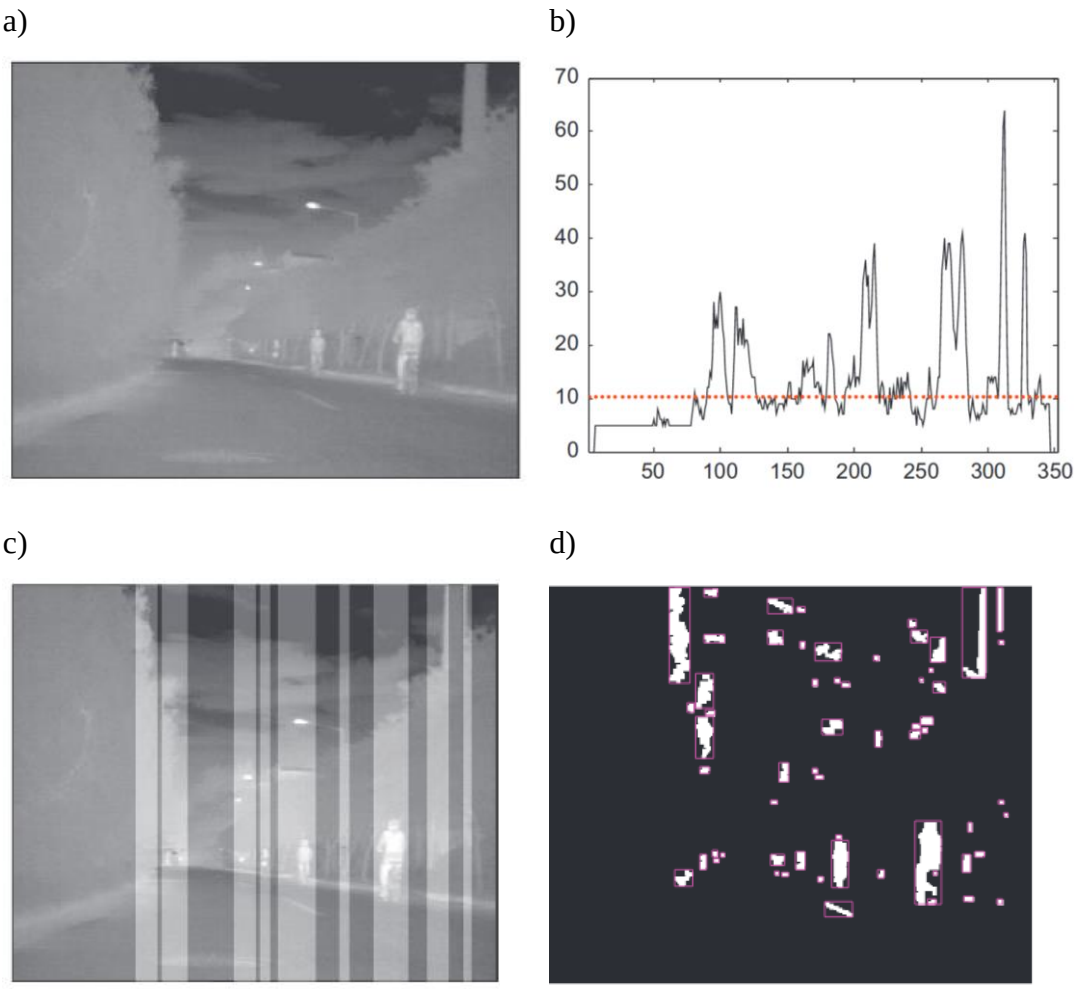


Figura 15. Generación de ROIs mediante la proyección vertical del gradiente: a) imagen de entrada, b) curva proyección vertical de gradiente, c) franjas verticales generadas, d) ROIs generados con umbral dual adaptativo aplicado solamente en las franjas verticales.

Fuente: (Liu, Zhuang, & Ma, 2013)

Para generar ROIs a peatones en imágenes infrarrojas también se utilizan algoritmos de agrupamiento como *mean shift*, el cual es utilizado en (Kim & Lee, 2013) o *c-means* el cual es utilizado en (John, Mita, Liu, & Qi, 2015), aquí se forman grupos de pixeles (conocidos como superpixeles) de similar intensidad que estén en

una misma vecindad, después se combinan los grupos que más alta intensidad tienen para formar los candidatos. La principal desventaja es que es relativamente costoso evaluar y asignar un grupo a cada pixel mientras que la ventaja radica en que se genera un número pequeño de ROIs por fotograma.

2.3.2 Clasificación

En reconocimiento de peatones basado en visión artificial existen dos enfoques, en el primero se generan características manualmente (handcrafted features) como HOG (Ge, Luo, & Tei, 2009), HAAR (Ge, Luo, & Tei, 2009), SURF (Besbes, Rogozan, Rus, Bensrhair, & Broggi, 2015), etc, luego estas características son alimentadas a algún algoritmo de clasificación como SVM ((O'Malley, Jones, & Glavin, 2010), (Liu, Zhuang, & Ma, 2013)), adaboost ((Davis & Keck, 2005), (Ge, Luo, & Tei, 2009)), redes neuronales artificiales (Lin, y otros, 2015), *random forest* (Lee, Ko, & Nam, 2016) mientras que en el segundo estas se aprenden automáticamente (feature learning) mediante algún algoritmo de aprendizaje profundo, comúnmente CNN ((Lee, Ko, & Nam, 2016), (Chen, 2015)).

2.3.2.1 Métodos de clasificación enfoque de características manuales

En aprendizaje automático la clasificación busca categorizar una nueva observación como perteneciente a una clase o categoría partiendo de un conjunto de datos que contienen observaciones o ejemplos cuya pertenencia se conoce previamente. Para realizar la clasificación se trabaja sobre los datos disponibles con el fin de extraer características representativas, informativas, no redundantes y relevantes que permitan representar o describir cada observación, luego las características son enviadas a un algoritmo de clasificación donde para nuestro caso, cada observación es categorizada como peatón o no peatón (Alpaydin, 2009). En resumen esta etapa se divide en dos partes: elección del algoritmo de clasificación y, método de extracción de características.

Uno de los primeros enfoques utilizado en clasificación de peatones en la noche se describe en (Fang, Yamada, Ninomiya, Horn, & Masaki, A shape-independent method for pedestrian detection with far-infrared images, 2004), aquí se describe al candidato utilizando el histograma, el inercial y el contraste de ese candidato, luego para la clasificación se mide la similaridad entre ese candidato y una imagen plantilla de peatón que es única y está generalizada estadísticamente. Para medir el rendimiento, aquí se valida el sistema sobre tres secuencias de video, una secuencia tomada en invierno, dos en verano: en un ambiente urbano y otra en un ambiente suburbano. El mejor resultado de clasificación se obtuvo en la secuencia de video de invierno con un AUC (por sus siglas en inglés Area Under the Curve) de 99%.

En (Davis & Keck, 2005) se describe a la imagen usando la magnitud del gradiente Sobel con un kernel de 3x3 para orientaciones de 0º, 90º, 45º y 135º y estas características son clasificadas mediante adaboost, este sistema se evalúa utilizando la base de datos OTCBVS la cual contiene 10 escenas, dando como resultado una sensibilidad promedio en las 10 escenas del 95%. En (Ge, Luo, & Tei, 2009) se representa a cada candidato inicialmente utilizando Haar wavelets, en base a estas características se entrena un pre clasificador adaboost, para finalmente pasar a un clasificador final (otro adaboost) entrenado utilizando características HOG (por sus siglas en inglés Histogram of Oriented Gradients) de los candidatos que pasen el pre clasificador, aquí se evalúa el rendimiento del detector con dos secuencias de video, en el primero se obtiene una tasa de detección de 96.1% y el segundo 90.6%, se muestra entonces que la combinación de HOG y HAAR se desempeña mejor que utilizar estos dos descriptores individualmente. Otro de los trabajos que también utiliza HOG para describir a sus candidatos es (O'Malley, Jones, & Glavin, 2010) aunque a diferencia del anterior aquí se utiliza como clasificador SVM (por sus siglas en inglés Support Vector Machine) obteniendo una tasa de detección del sistema de 96% en una secuencia de video de 834 fotogramas. En (Liu, Zhuang, & Ma, 2013) se utiliza una variante de HOG conocida como EWHOG, la idea de este descriptor es que los componentes de HOG extraídos de las regiones periféricas locales de los peatones pueden ponderarse más que los de las regiones internas de los peatones, aquí se usa como clasificador una estructura de tres ramas SVM, aquí se utiliza 4 secuencias de

video para evaluar el sistema, 2 de invierno y 2 de verano obteniendo una tasa de detección promedio de 96.7%.

En (Olmeda, de la Escalera, & Armingol, Contrast invariant features for human detection in far infrared images, 2012) se aplica la congruencia de fase a cada ROI para reducir los efectos causados por el cambio de iluminación, estas características generadas con la congruencia de fase se las entrena con SVM.

Otros enfoques combinan descriptores uniendo características originadas de varios métodos independientes como en (Ma, Chen, & Chen, 2011) y (Enzweiler & Gavrila, 2011).

2.3.2.2 Métodos basados en aprendizaje profundo

Tal como se define en (Guo, y otros, 2016) y (Deng, 2014), aprendizaje profundo es un subcampo del aprendizaje automático que aprende características de alto nivel utilizando varias capas de etapas de procesamiento de información en arquitecturas jerárquicas que son explotadas en aprendizaje de características con entrenamiento no supervisado especialmente (unsupervised feature learning) y en reconocimiento/clasificación de patrones.

La **Figura 16** muestra la clasificación de los métodos de aprendizaje profundo con los trabajos más importantes de cada categoría, de ellas, solamente los métodos basados en CNN (por sus siglas en inglés, Convolutional Neural Networks), RBM (por sus siglas en inglés, Restricted Boltzmann Machines) y *autoencoder* tienen la propiedad de aprender características automáticamente (feature learning) por lo cual son usados en aplicaciones de visión por computador, aunque CNN es el más conocido, usado y desarrollado teóricamente (Jain, 2016).

Las CNN son un tipo especial de redes neuronales de la clase *feed-forward* que están inspiradas en la visión animal y han sido diseñadas para trabajar sobre datos bidimensionales, tales como imágenes y videos (Arel, Rose, & Karnowski, 2010). En

CNN existe una íntima relación entre las capas y la información espacial 2D por lo que es comúnmente usado en aplicaciones de visión por computadora ((Karpathy, s.f.), (An Intuitive Explanation of Convolutional Neural Networks, 2016), (Theano, 2013)).

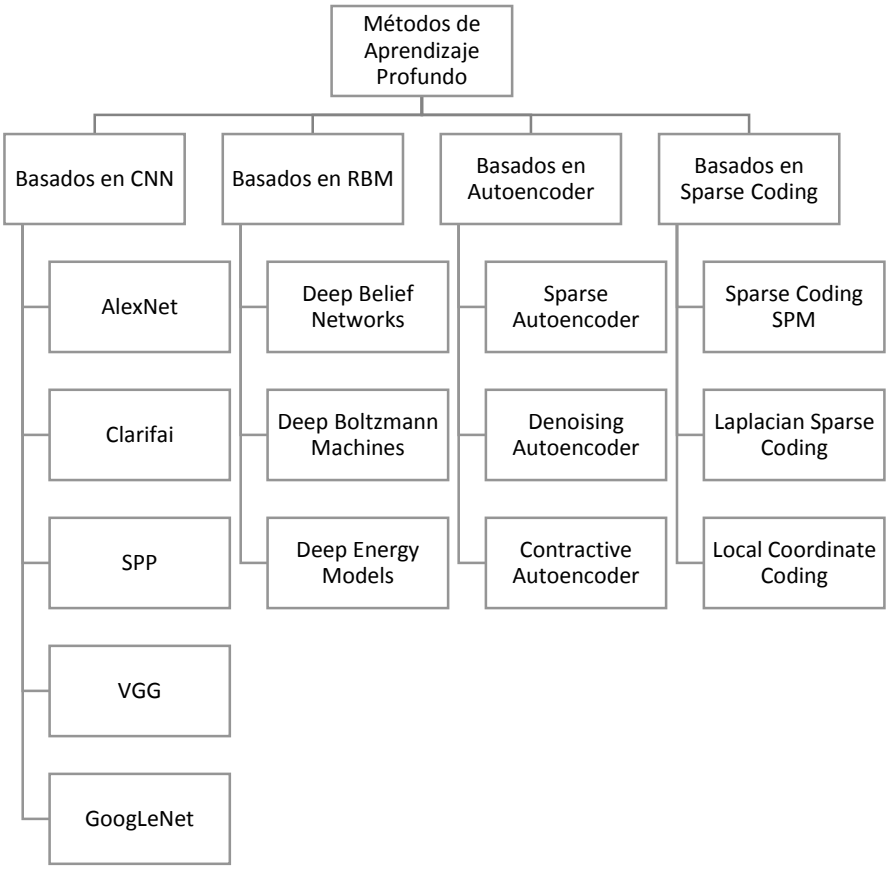


Figura 16. Métodos de Aprendizaje Profundo con sus aplicaciones más importantes. Fuente: (Guo, y otros, 2016)

Generalmente una arquitectura CNN ((Krizhevsky, Sutskever, & Hinton, 2012), (Szegedy, y otros, 2015), (Simonyan & Zisserman, 2014), (Sermanet, y otros, 2013)) consta de tres capas principales: capas convolucionales, capas de agrupación (pooling layers) y capas completamente conectadas entre sí (fully-connected layers). Esta última no es nada más que un MLP (An Intuitive Explanation of Convolutional Neural Networks, 2016), la **Figura 17** muestra un ejemplo de reconocimiento.

Dependiendo de qué tan profundo se quiera llegar, una CNN se forma por la combinación jerárquica de estas tres capas principales, desgraciadamente no existe un método que permite determinar la profundidad de una red CNN esto depende directamente de la experiencia e intuición del diseñador (Wurfl, 2016) (Baker, Gupta, Naik, & Raskar, 2016).

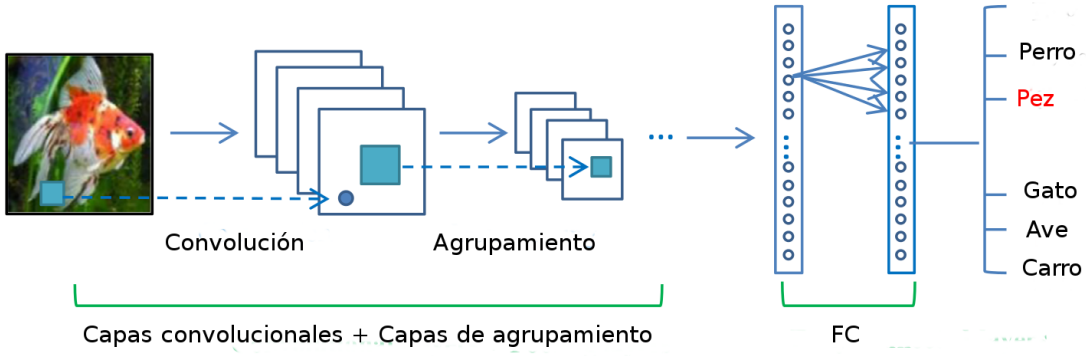


Figura 17. Arquitectura general de una red CNN, donde se observa la imagen de entrada, la capas convolucionales, las capas de agrupamiento y las capas FC.
Fuente: (Guo, y otros, 2016)

Las CNN se utilizan en clasificación de objetos y en reconocimiento de imágenes, debido a que permiten definir la generación de características y la clasificación per se como parte de una sola etapa, es decir, las características se aprenden conjuntamente con los parámetros de clasificación. Las CNN también ofrecen otras ventajas como: conectividad local, compartición de parámetros entre unidades e invariancia en la traslación.

Cuando se tiene entradas de alta dimensionalidad como imágenes es impráctico conectar una unidad con todas las unidades de la capa anterior (como en MLP) en vez de eso en CNN cada unidad se conecta con una región local de la capa anterior cuya extensión espacial es un hiperparametro llamado campo receptivo (receptive field) y es equivalente al tamaño del filtro o kernel utilizado en esa capa convolucional (Karpathy, s.f.). Esto significa que se requiere menos memoria computacional para el modelo y menos operaciones son realizadas para calcular la activación.

En una CNN cuando se desplaza un kernel sobre la imagen de entrada cada miembro del kernel (los parámetros) recorre toda la imagen, es decir a diferencia de los MLP

donde cada parámetro se utiliza una sola vez y es independiente para cada conexión, en CNN algunos miembros del kernel se comparten entre unidades. Esto es beneficioso debido a que en lugar de aprender parámetros por cada unidad, separados e independientes entre sí, se aprende un conjunto limitado y relacionado de cierta forma. Otra ventaja de las CNN es que gracias a las capas de agrupamiento se añade invariancia a pequeñas traslaciones en la entrada, por ejemplo si se desplaza la entrada la mayoría de unidades de salida de la capa de agrupamiento siguiente no cambian, esto debido a que se agrupa todas las posibles traslaciones (dentro de cierto campo receptivo) dentro de un solo pixel con lo cual la activación sigue existiendo independientemente de donde se encuentre la característica a detectar (Goodfellow, Bengio, & Courville, 2015).

CNN se utiliza en la detección de objetos en (Girshick, Donahue, Darrell, & Malik, 2014), donde se definió con el nombre de redes neuronales convolucionales basado en regiones (R-CNN por sus siglas en inglés Region based Convolutional Neural Network) aquí se propone primeramente generar mediante algún método de generación de regiones de interés candidatos de varias clases, para luego con cada candidato generado fijar su resolución y producir características mediante la red convolucional, para luego clasificar estas características mediante varios SVMs lineales entrenados específicamente para cada clase. Adicionalmente se tiene en la última capa un algoritmo de regresión (llamado en inglés bounding box regression) el cual predice los desplazamientos que puedan tener las regiones de interés a partir del tiempo en el cual se generó, este desplazamiento se añade a las coordenadas inicialmente detectadas por el método de generación de candidatos. En (Girshick, Donahue, Darrell, & Malik, 2014) usan específicamente SVM como clasificador pero se puede utilizar otros algoritmos de clasificación como MLP o regresión logística.

En (Chen, 2015) y (Tome, y otros, 2016) se utiliza R-CNN como detector de peatones al limitar los candidatos generados, aquí se trabaja sobre imágenes a color, en contraste con (John, Mita, Liu, & Qi, 2015) que trabaja con imágenes infrarrojas lejanas.

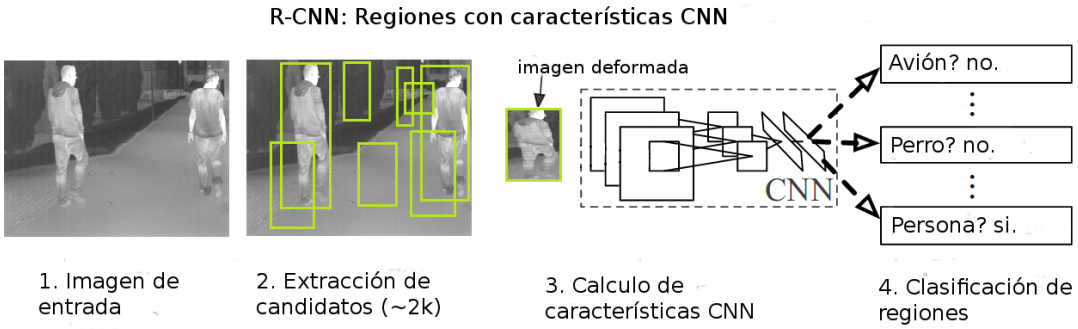


Figura 18. Vista general de la detección de objetos con R-CNN modificado al caso de detección de peatones en la noche.
Fuente: (Girshick, Donahue, Darrell, & Malik, 2014)

El problema con los métodos basados en R-CNN es que utilizan arquitecturas profundas para generar características, con lo cual, para su entrenamiento se necesita mucha memoria RAM, tal como se menciona en (Girshick, Fast r-cnn, 2015) toma 2.5 días GPU entrenar una arquitectura VGG16 (Simonyan & Zisserman, 2014) usando 5000 imágenes de la base de datos de entrenamiento VOC07 (Everingham, Van-Gool, Williams, Winn, & Zisserman, 2017). El segundo problema surge cuando se tiene un vasto conjunto de candidatos generados ya que cada uno de ellos debe pasar por toda la arquitectura uno a la vez tal como se observa en la **Figura 19**, lo que consume demasiados recursos computacionales solo para detectar un candidato, en (Girshick, Fast r-cnn, 2015) se dice que la detección toma 47 segundos por fotograma usando aproximadamente 2000 candidatos utilizando la arquitectura VGG16.

En las arquitecturas usadas en R-CNN las capas FC (por sus siglas en inglés Fully Connected) tiene un número de unidades fijo que no puede cambiar, debido a esta razón, los candidatos generados entran a la arquitectura con un tamaño de resolución fijado a priori el cual genera siempre el número de características que se desea, tal como se muestra en la **Figura 19**. Fijar la resolución de un candidato puede generar distorsión geométrica (**Figura 20**) y debido a la gran variabilidad de escalas y relaciones de aspecto que puede presentar un candidato, esto puede bajar el rendimiento de clasificación. Finalmente debido a la profundidad (número de capas)

que se tiene en las arquitecturas disponibles en el estado del arte, para entrenar una de estas arquitecturas se necesita un conjunto de entrenamiento bastante grande.

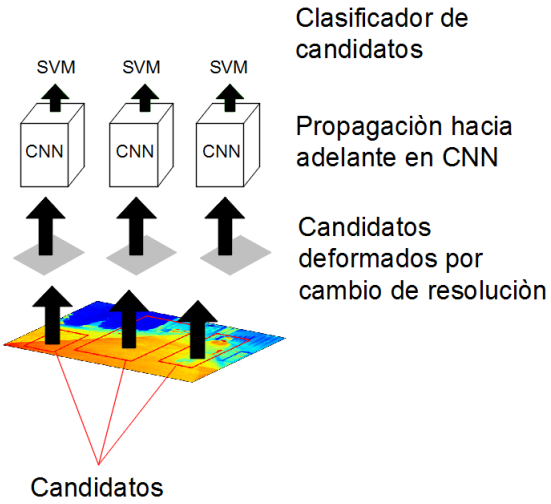


Figura 19. Enfoque R-CNN para la detección de objetos.
Fuente: (Girshick, Fast R-CNN, 2015).

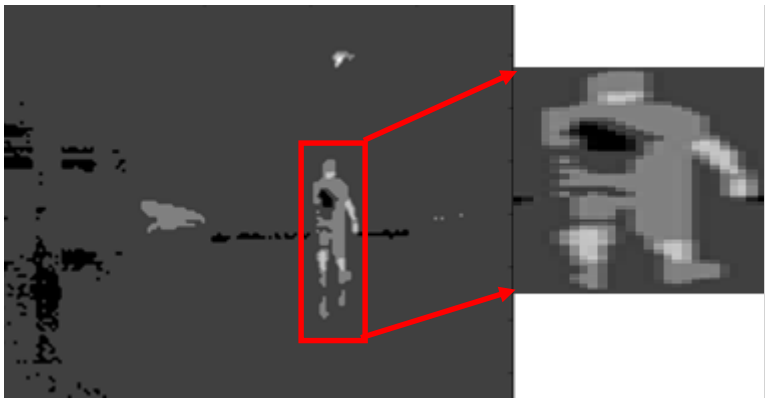


Figura 20. Ejemplo de la distorsión de la imagen al escoger un candidato y cambiar su resolución.
Fuente: (He, Zhang, Ren, & Sun, 2014).

En (He, Zhang, Ren, & Sun, 2014) se resuelve el problema de cambiar la resolución de los candidatos en la entrada de la arquitectura mediante la invención de una nueva capa a la cual denominaron SPP (por sus siglas en inglés Spatial Pyramid Pooling), esta capa es parecida a la capa de agrupamiento (pooling) excepto que se tiene híperparámetros variables que permiten generar características constantes que se

aplican a varios niveles formando una pirámide tal como se muestra en la **Figura 21**, después las características generadas por cada nivel de la pirámide se concatenan entre sí, en resumen permite obtener un número constante de características independientemente de la resolución del candidato. En (Lee, Ko, & Nam, 2016) se presenta una arquitectura con dos capas convolucionales que detecta peatones en ciertas posiciones como corriendo, caminando, parado, sentado, de perfil y acostado. Esta arquitectura usa la capa SPP sobre la segunda capa convolucional para generar características y como clasificador el algoritmo *random forest* (Criminisi & Shotton, 2013).

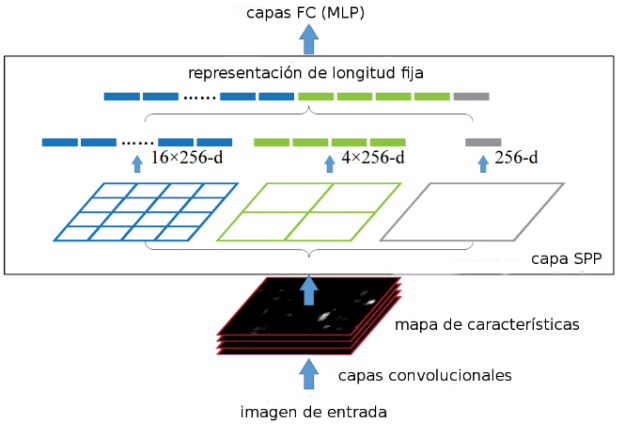


Figura 21. Funcionamiento de la capa SPP.
Fuente: (He, Zhang, Ren, & Sun, 2014)

Las arquitecturas que usan la capa SPP (SPPnet) tienen la desventaja de que ocupan mucho tiempo de entrenamiento por la cantidad de memoria requerida (Girshick, Fast R-CNN, 2015). En la propuesta *fast* R-CNN (Girshick, Fast r-cnn, 2015) con una implementación más simple se logró mejorar los tiempos de entrenamiento al utilizar un solo nivel de la pirámide en la capa SPP a la que se denominó capa roi *pooling*, esta capa proyecta cada candidato dentro del mapa de activaciones final para luego generar un volumen de datos cuyo tamaño es independiente del tamaño del candidato tal como se muestra en la **Figura 22**, esta técnica se describe a profundidad en (Girshick, Fast r-cnn, 2015). *Fast* R-CNN se utiliza en la detección de peatones en imágenes RGB en (Li, y otros, 2015) donde se tiene dos subredes que permiten detectar peatones de escala pequeña y grande.

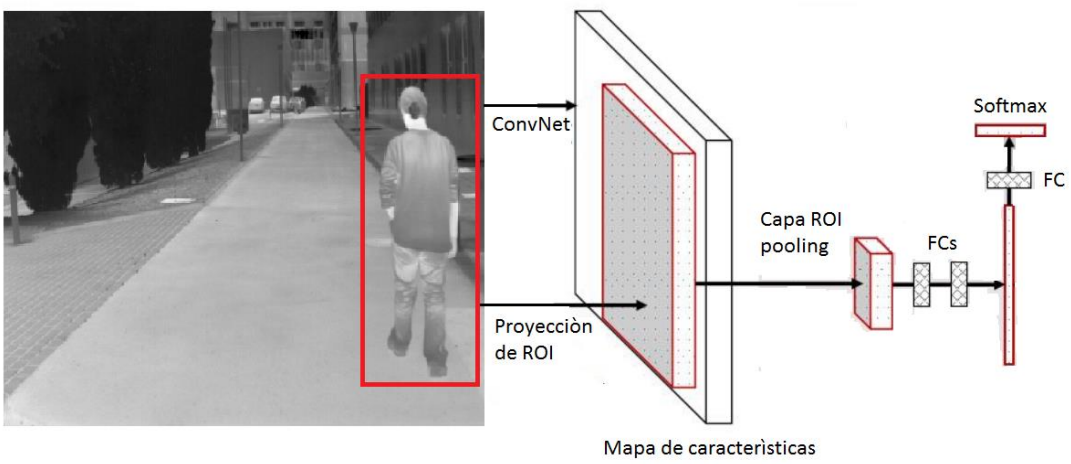


Figura 22. Vista general de la arquitectura *fast* R-CNN modificado al caso de detección de peatones en la noche.
Fuente: (Girshick, Fast R-CNN, 2015)

2.3.2.2.1 Aplicación de arquitecturas CNN a un problema de reconocimiento especifico

No existe un método exacto que permita determinar que arquitectura se desempeña mejor sobre algún problema de reconocimiento, ni tampoco existe un método que permita conocer los hiperparámetros de una CNN, como el número de capas, el número de filtros por capa, el tamaño del filtro, el tamaño del *stride*, entre otros.
En general existen tres enfoques para afrontar este problema:

- El primero y el más común es reutilizar alguna arquitectura disponible como LeNet-5 (LeCun, Bottou, Bengio, & Haffner, 1998), AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), ZF net (Zeiler & Fergus, 2014), o VG-16 net (Simonyan & Zisserman, 2014) para sintonizar o reentrenar utilizando nuestros datos partiendo de los parámetros que ya fueron entrenados anteriormente. El problema con este enfoque es que la mayoría de arquitecturas disponibles son de tipo profundas, es decir tienen muchas capas que necesitan sintonización y para ciertas aplicaciones podría significar un desperdicio de procesamiento en tareas que no requieran tanta computación o en ciertos casos debido a la cantidad de

procesamiento simplemente no se podría resolver en una computadora convencional (Karpathy, s.f.).

- El segundo enfoque utiliza las mismas arquitecturas pre entrenadas mencionadas anteriormente, pero en este caso utiliza la red convolucional como extractor de características y solo se entrena el clasificador sobre estas características (Karpathy, s.f.).
- El tercer enfoque es el que se describe en (Weimer, Scholz-Reiter, & Shpitalni, 2016), para ciertas aplicaciones en la cual se tiene cierta certeza de convergencia con una arquitectura pequeña, se puede escoger manualmente los hiperparámetros de la red y experimentalmente variar uno de estos parámetros y determinar la mejor arquitectura teniendo en cuenta el tiempo de procesamiento y la exactitud, o también se puede partir de problemas similares y mediante experimentación llegar a la arquitectura adecuada para nuestro problema. El problema de este enfoque es que se requiere mucha experiencia del diseñador y algunas veces no se logra una convergencia rápida con una arquitectura pequeña y se podría terminar entrenando una arquitectura profunda cuyo entrenamiento desde cero podría llevar varios días o incluso semanas. Debido a evidencia encontrada en el estado del arte (Lee, Ko, & Nam, 2016) esta aplicación puede converger con una arquitectura pequeña de dos capas convolucionales es por esto que en este trabajo se va a utilizar este último enfoque.

2.4 Conclusiones

En la detección de peatones en la noche para la generación de ROIs inicialmente se comenzó con el método de la ventana deslizante, después debido principalmente al alto costo computacional, se pasó a la generación de candidatos basado en la segmentación de la imagen mediante la aplicación de umbrales globales para todo el fotograma, luego debido a la uniformidad de intensidades que se puede tener entre el peatón y el fondo, se cambió el enfoque y se utilizó umbrales adaptativos que se aplican por pixel para finalizar con métodos basados en algoritmos de agrupamiento que forman superpixeles de similar intensidad. De los métodos mencionados se puede

decir que sus parámetros en la mayoría de los casos necesitan ser sintonizados manualmente para cada ambiente específicamente.

En el reconocimiento de peatones se puede decir que los métodos basados en CNN aprenden parámetros automáticamente optimizados para un problema específico de clasificación a diferencia de los métodos de clasificación con características manuales los cuales requieren un alto nivel de experticia por parte del diseñador.
Fast R-CNN mejora grandemente los tiempos de procesamiento en entrenamiento (9 veces menos comparado con R-CNN medido en una arquitectura VGG16) y ejecución (146 veces menos comparado con R-CNN medido en una arquitectura VGG16) de R-CNN gracias a la capa roi-*pooling* la cual selecciona cada candidato del último mapa de características para su clasificación.

CAPÍTULO III
3. MÉTODO DE GENERACIÓN DE CANDIDATOS A PEATONES EN IMÁGENES INFRAROJAS

3.1 Introducción

La generación de candidatos se define como la búsqueda de regiones con alta probabilidad de pertenecer a cierta clase, esta pertenencia se puede definir concretamente con algún algoritmo de clasificación. Para la selección del candidato, se puede escoger solo su silueta o como en la mayoría de los casos se puede escoger un rectángulo delimitador (BB, por sus siglas en inglés Bounding Box) que incluye parte del fondo (Olmeda, Pedestrian Detection in Infrared Images, 2014).

Para este trabajo se consideró dos métodos de generación de candidatos a peatones, el primer método está orientado a detectar peatones que sean fácilmente reconocibles en el fotograma, es decir peatones que resalten en el fotograma al presentar una alta intensidad contrastado con un fondo de muy baja intensidad, el segundo método está orientado a detectar peatones que tenga poco contraste con el fondo, es decir que presenten similares intensidades entre sí.

Finalmente por cada método de generación de ROIs se genera una lista de candidatos las cuales son concatenadas y alimentadas a la etapa de clasificación como se observa en **Figura 23**.

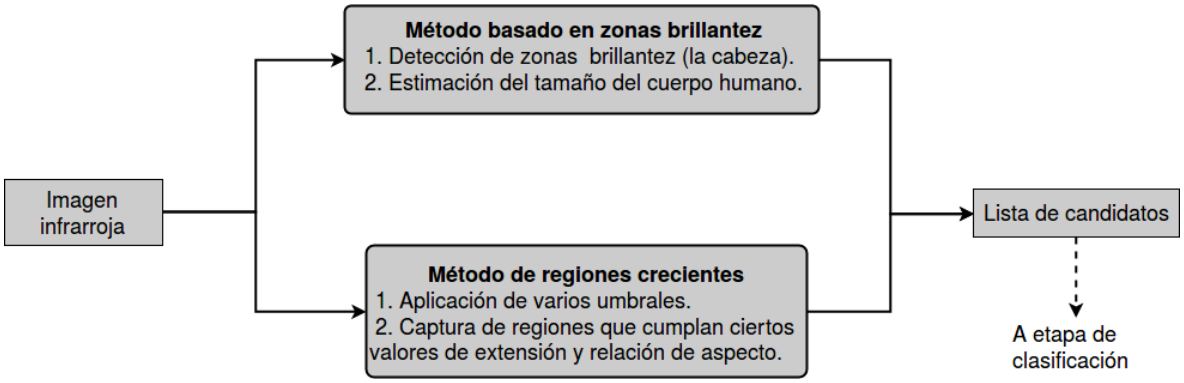


Figura 23. Esquema global para la generación de ROIs sobre imágenes en el infrarrojo.

3.2 Generación de candidatos a peatones en imágenes con bajo contraste entre el peatón y el fondo

Con este método se plantea cubrir el espectro de candidatos cuyo contraste con el fondo es bajo. En la **Figura 24** se presenta el histograma de la imagen donde se observa el bajo contraste entre el fondo y el peatón.

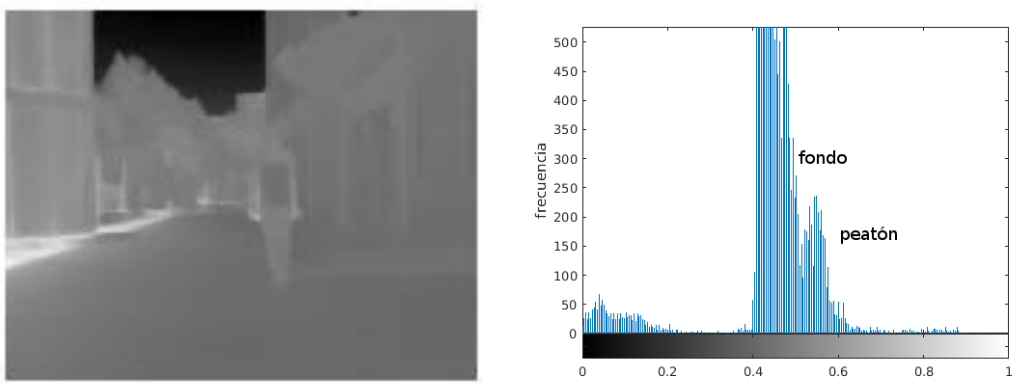


Figura 24. Ejemplo de fotograma con bajo contraste entre el peatón y el fondo. Fuente: (Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013).

3.2.1 Generación de cuerpos en imágenes en el infrarrojo

Un cuerpo es una región de pixeles que comparten una misma característica, como por ejemplo el mismo color o el mismo valor de intensidad en imágenes en escala de grises. Existen muchos enfoques para detectar estos cuerpos como por ejemplo los basados

en el Laplaciano del Gaussiano (LoG por sus siglas en ingles Laplacian of Gaussian), o el enfoque conocido como Diferencia del Gaussiano (DoG por sus siglas en inglés Difference of Gaussians) el cual pretende aproximar el LoG mediante diferencias de gaussianos a diferentes escalas. Sin embargo en este trabajo se utiliza *fast-hessian detector* el cual es una versión rápida del conocido enfoque basado en el determinante del hessiano (DoH por sus siglas en inglés Determinant of Hessian). Este detector se explica con detalle en (Bay, Ess, Tuytelaars, & Van Gool, 2008), donde se lo conoce como *SURF: Speeded Up Robust Features*.

Para el caso de detección de peatones en la noche se aplica este método para detectar regiones de alta intensidad en una vecindad de baja intensidad que probablemente representaría la cabeza del peatón. A continuación se muestran los cuerpos detectados en la segunda octava, se puede ver que desde la tercera capa se comienza a tener detecciones como la cabeza y los brazos. La cantidad de detecciones depende del número de octavas y del número de capas por octava.

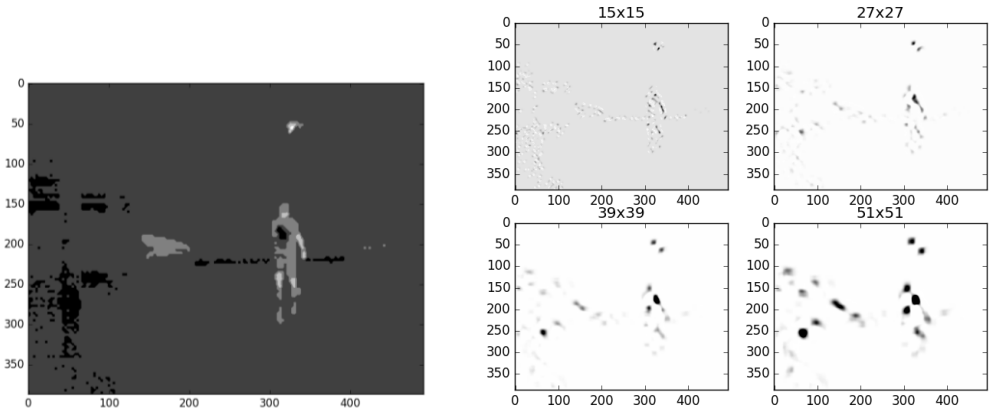


Figura 25. Detección de cuerpos para la segunda octava.
Fuente: (Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013).

3.2.2 Definición de candidatos

El detector de cuerpos basado en el determinante del hessiano detecta regiones con alta intensidad por lo cual puede ser utilizado como detector de cabezas en el caso ideal ya que generalmente son las regiones de la imagen que más intensidad presentan. El

algoritmo descrito en (Bay, Ess, Tuytelaars, & Van Gool, 2008) provee además de la localización del cuerpo, el diámetro de la región el cual está dado por su escala σ, el cual podría interpretarse como el diámetro de la cabeza.

Suponiendo que se tiene una detección correcta de una cabeza, el cuadro delimitador del candidato puede ser formado conociendo que proporción del cuerpo representa una cabeza. Según (Proporciones del cuerpo humano, 2013) el canon de cuerpo humano adulto tiene aproximadamente una altura de 8 cabezas y un ancho de 2 cabezas.

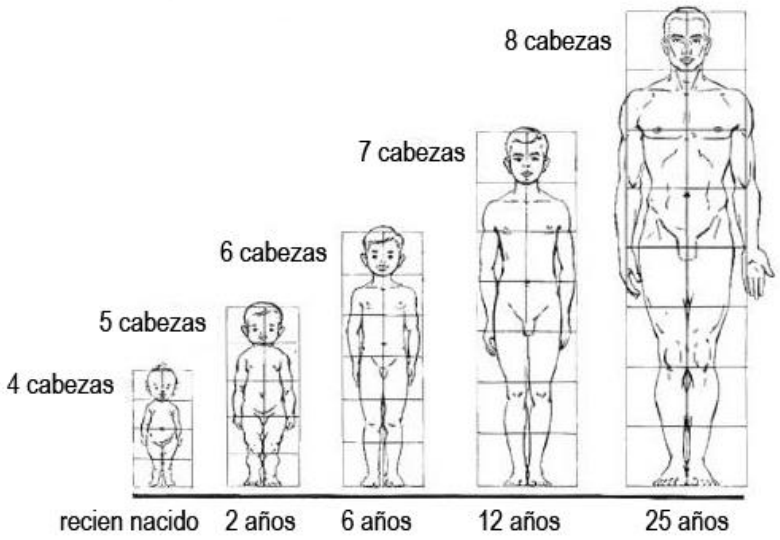


Figura 26. Proporciones del cuerpo humano.
Fuente: (Proporciones del cuerpo humano, 2013)

Entonces con la ubicación del centro y diámetro de la cabeza, se puede definir el cuadro delimitador según las siguientes ecuaciones:

$$x = c_x - diametro \tag{8}$$

$$y = c_y - diametro/2 \tag{9}$$

$$w = 2 * diametro \quad (10)$$

$$h = 8 * diametro \quad (11)$$

En situaciones donde se tienen peatones en pequeña escala o para detecciones de puntos que no sean la cabeza, estas ecuaciones no enmarcan correctamente al peatón. Por lo cual para detectar peatones en las condiciones mencionadas anteriormente se pretende cumplir con la mayoría de los casos, para esto se va a trabajar con las opciones que se cubren en la **Figura 26** y con redundancia adicional añadida en la implementación.

3.3 Generación de candidatos a peatones en imágenes con alto contraste entre el peatón y el fondo

Con este método se plantea cubrir el espectro de candidatos cuyo contraste con el fondo es alto, es decir el peatón y el fondo se encuentran bien separados. La Figura 27 muestra el histograma que ejemplifica este caso. En estas circunstancias es posible aplicar un método basado en la selección de umbrales para segmentar y generar los candidatos.

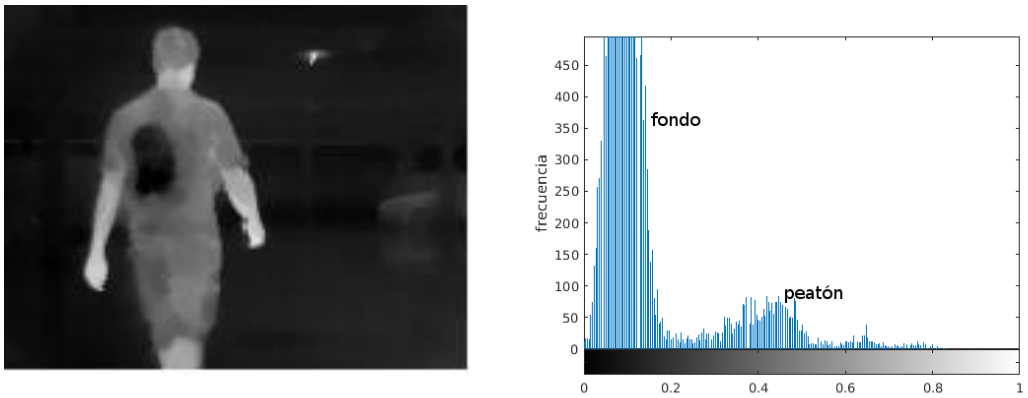


Figura 27. Ejemplo de fotograma con alto contraste entre el peatón y el fondo.
Fuente: (Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013)

Para detectar los candidatos se consideró el método de regiones crecientes descrito en (O'Malley, Jones, & Glavin, 2010), el cual consiste en aplicar varios

valores de umbrales para generar un conjunto de imágenes binarias, correspondientes a cada umbral y capturar los segmentos de cada fotograma que cumplan con valores de extensión y relación de aspecto fijados a priori. La **Figura 28** describe el algoritmo utilizado.

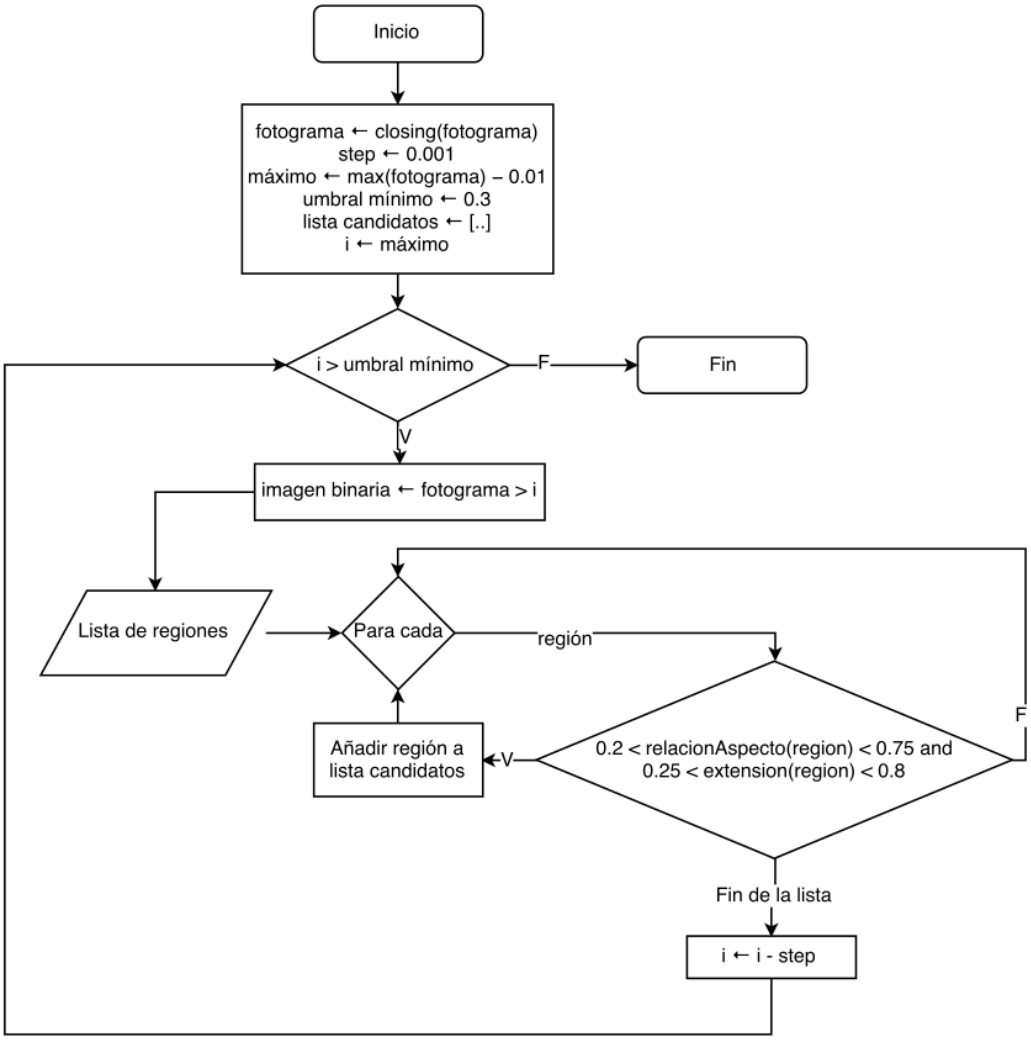


Figura 28. Esquema del algoritmo para la generación de candidatos a peatones. Fuente: (O'Malley, Jones, & Glavin, 2010)

3.4 Conclusiones

En este capítulo se han introducido dos técnicas para la generación de ROIs, la primera consiste en buscar las zonas más brillantes de la imagen, que corresponde con alta probabilidad a la cabeza del peatón, a partir de lo cual se construye las ROIs. Esta técnica es aplicable cuando no existe suficiente contraste entre los peatones y el fondo.

La segunda está constituida por el método de regiones crecientes, usando distintos umbrales, y es aplicable al caso cuando el fondo y los peatones presentan un contraste importante.

De esta manera se cubre un amplio espectro de casos en la generación de ROIs, en condiciones variables de iluminación durante la noche.

CAPÍTULO IV

4. RECONOCIMIENTO DE PEATONES EN EL INFRARROJO MEDIANTE *FAST* R-CNN

4.1 Introducción

En este capítulo se presenta el diseño de una nueva arquitectura basada en *fast* R-CNN para la clasificación de peatones en la noche. Luego, se describe el método de entrenamiento y finalmente se describe la refinación de la función de decisión.

4.2 Diseño de la arquitectura *fast* R-CNN para la clasificación de peatones en la noche

Tal como se mencionó en el estado del arte, en la sección 2.3.2.2.1 para este trabajo se va a utilizar el tercer enfoque para el diseño de la arquitectura, se van a probar experimentalmente tres arquitecturas *fast* R-CNN para la clasificación de peatones. Inicialmente se determina desde que arquitectura base se quiere partir y se varía un solo parámetro cada vez con el fin de comparar el rendimiento de estas arquitecturas. Tal como se puede observar en la **Tabla 5**, la primera arquitectura consta de una capa convolucional seguida de un clasificador de dos capas, la segunda arquitectura se diferencia de la primera en que se añade una capa convolucional mas y se mantiene el clasificador tal y como estaba. En la tercera arquitectura se modifica la capa escondida del clasificador y se mantienen las capas convolucionales de la red anterior.

Para tener un mejor entendimiento de la arquitectura es necesario conocer el significado de la terminología usada en la **Tabla 5.**: conv5-Nk1, conv hace referencia a una capa convolucional, 5 hace referencia al tamaño de filtro a utilizar para esa capa, en este caso de 5x5 y Nkx hace referencia al número de filtros a utilizarse en esa capa y es un valor que se asigna arbitrariamente para comenzar la experimentación.
ReLU hace referencia a la función de activación llamada *rectified linear unit* (ReLU) descrita por la siguiente ecuación:

$$f(x) = \max(0, x) \tag{12}$$

La capa max-pooling es una capa de reducción de dimensionalidad o *downsampling* la cual permite reducir espacialmente el tamaño del resultado de las capas convolucionales a través de alguna operación, como el valor máximo o el promedio de una región de pixeles (ver el anexo **A.3.2 Capas de agrupamiento**).

La capa roi *pooling* genera una salida de tamaño constante por cada candidato, sin importar el tamaño de este. Esta capa proyecta la región de interés dentro de la última capa convolucional y extrae una región equivalente y luego aplica max-pool con hiperparámetros variables dependiendo del tamaño del candidato esto con el fin de generar una salida constante que pueda ser alimentada al clasificador final (Girshick, Fast r-cnn, 2015).

FC (por sus siglas en inglés Fully Connected layer) hace referencia a una sola capa de MLP (por sus siglas en inglés, MultiLayer Perceptron), específicamente es la capa escondida del clasificador que a su vez se encuentra seguido de una función de activación ReLU. El número que acompaña a FC-x especifica el número de unidades en esa capa.

En la salida no se aplica función de activación, más bien se aplica directamente regresión logística con dos salidas. En la literatura revisada generalmente en CNNs se utiliza SVMs lineales o Softmax (para clasificación multiclase) como clasificadores, en este trabajo se utiliza como clasificador regresión logística (softmax para el caso en el cual se tienen dos clases) ya que ligeramente supera en rendimiento a SVM en arquitecturas *fast* R-CNN según el análisis mostrado en (Girshick, Fast r-cnn, 2015).

Tabla 5.
Tabla de arquitecturas propuestas para la detección de peatones sobre imágenes en el infrarrojo.

	Base	**Arquitectura experimental**	**Arquitectura propuesta**
Capa convolucional con kernel 5x5	conv5-Nk1	conv5-Nk1	conv5-Nk1
Función de activación	ReLU	ReLU	ReLU
Capa de agrupamiento		max-pooling	max-pooling
Capa convolucional con kernel 3x3		conv3-Nk2	conv3-Nk2
Función de activación		ReLU	ReLU
Capa roi-pooling	roi-pooling	roi-pooling	roi-pooling
Capa FC	FC-500	FC-500	FC-1024
Función de activación	ReLU	ReLU	ReLU
Capa FC	FC-2	FC-2	FC-2
Clasificador	regresión logística	regresión logística	regresión logística

Tabla 6.
Rendimiento de las arquitecturas de la tabla anterior sobre las bases de datos LSI y CV-09 (solo clasificación).

Arquitectura	**Exactitud [μ = 0.5]**	**Dimensión de características**	**Tiempo por fotograma para 100 candidatos [segundos]**
Base	0.70	500	0.046

Continúa

Arquitectura experimental	0.78	500	0.089
Arquitectura propuesta	0.95	1024	0.135

Para comenzar a experimentar con las arquitecturas propuestas se asignó Nk1 un valor de 50 y Nk2 un valor de 100 obteniéndose como resultado la **Tabla 6**.

En esta tabla la exactitud se obtiene al comparar los candidatos que se alimentan a la arquitectura con los candidatos obtenidos a la salida de esta. Las experimentaciones con las arquitecturas se hicieron usando un umbral de decisión de 0.5, es decir si la unidad de salida correspondiente a peatones es mayor que 0.5 se consideró a ese candidato un peatón.

La primera arquitectura muestra un rendimiento regular con una exactitud de clasificación de 70% debido seguramente al pobre modelo elegido en la arquitectura, el cual no permite de alguna manera generar características lo suficiente discriminatorias entre un ejemplo positivo y otro negativo, esto se conoce en aprendizaje automático como *underfitting*. En la segunda arquitectura se obtiene 78% de exactitud al aumentar una capa convolucional, lo cual al no tener una mejora significativa sugiere que se debería mejorar de alguna forma el clasificador. En la última arquitectura se aumentaron unidades en la capa escondida, con lo cual, al tener un modelo más complejo se obtuvo un mejor rendimiento.

4.3 Descripción de arquitectura *fast* R-CNN para la detección de peatones en la noche sobre imágenes en el infrarrojo

En la **Figura 29** se muestra el esquema global del sistema utilizado, donde se destaca principalmente que la extracción de características de todos los candidatos se hace una sola vez y al mismo tiempo y la clasificación se hace por candidato una vez que las características están calculadas.

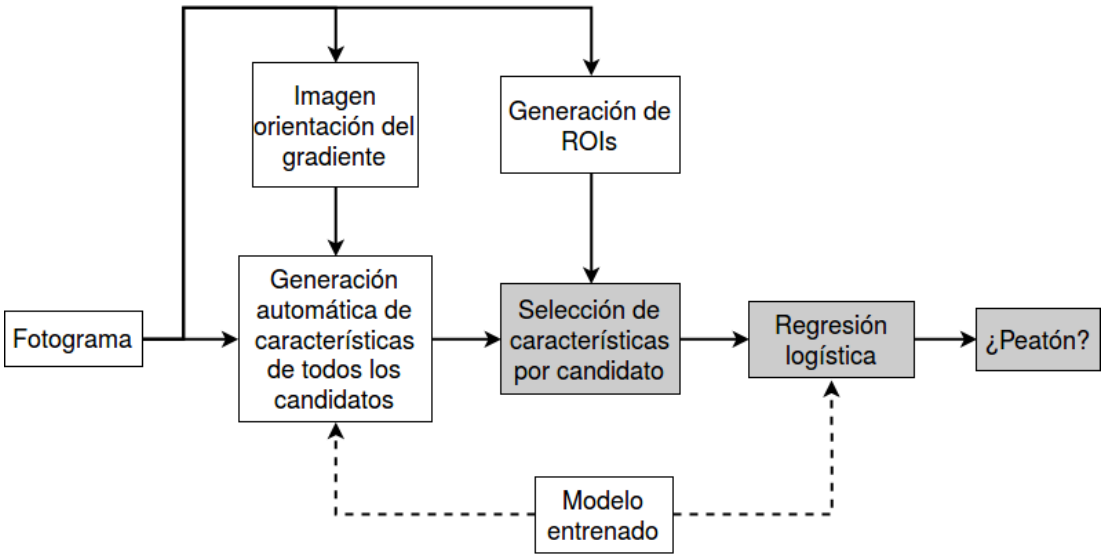


Figura 29. Esquema global del sistema de detección de peatones en la noche, los cuadros grises se ejecutan por candidato.

En la **Figura 30** se muestra la arquitectura *fast* R-CNN implementada en este trabajo, aquí se muestran las activaciones que se tienen en cada capa indicadas según la siguiente notación: cantidad N x canal K x alto H x ancho W. Según esto se puede determinar que en la arquitectura ingresa una imagen de dos canales, el primer canal es el fotograma en escala de grises, el segundo canal es la orientación del gradiente, dicho gradiente brinda importante información sobre los bordes del candidato.

En la primera capa se utilizan 50 filtros de 5x5, en la segunda capa se utilizan 100 filtros de 3x3, dichos filtros fueron aprendidos durante el entrenamiento. En la **Figura 32** se observan las disposiciones de los filtros de la arquitectura.

En el clasificador se utiliza una capa escondida de 1024 unidades, seguida de una capa de salida de 2 unidades, y sobre esta capa se utiliza regresión logística para representar la salida en forma de distribución de probabilidad sobre las dos categorías o clases posibles peatón y no peatón.

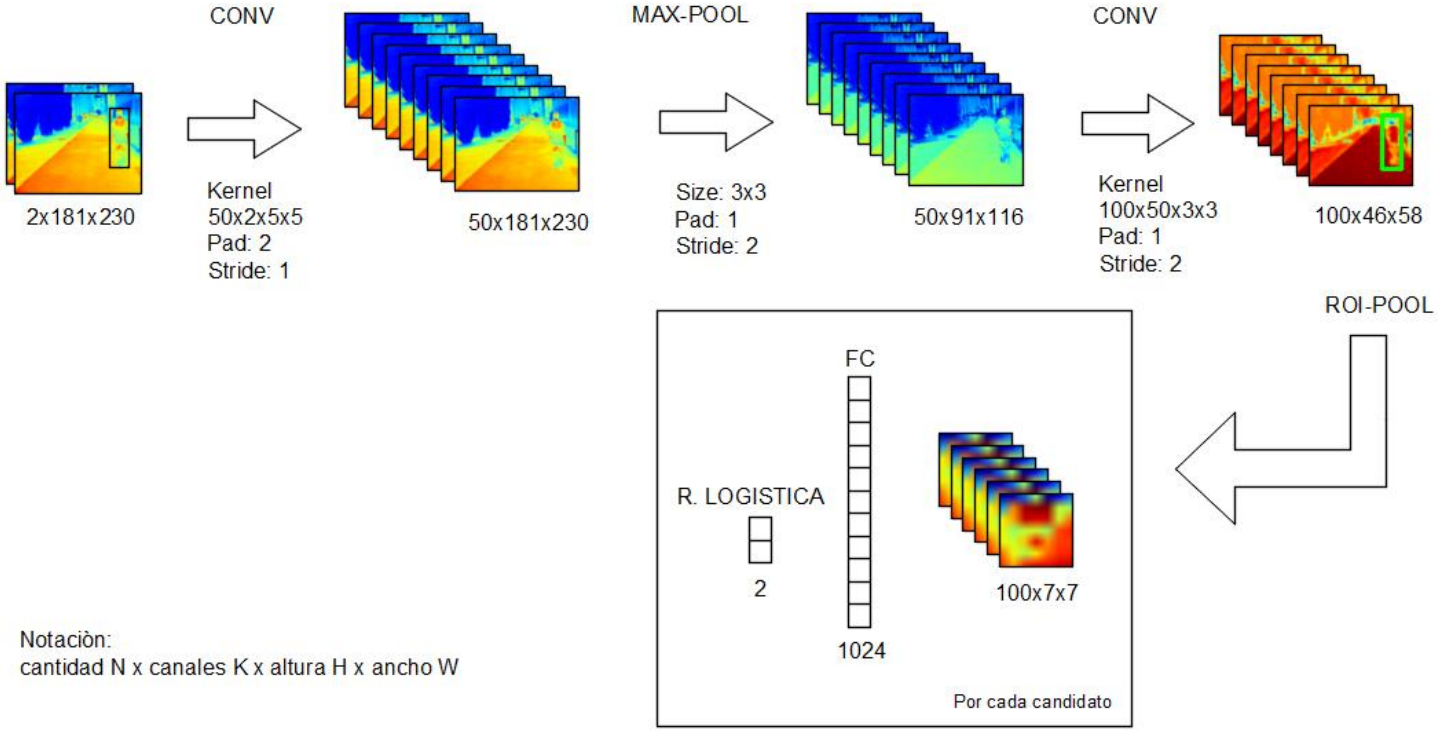


Figura 30. Arquitectura *fast* R-CNN para la clasificación de peatones por la noche sobre imágenes en el infrarrojo.

En este trabajo se propone una arquitectura de dos capas convolucionales pero a diferencia de los trabajos basados en R-CNN no se ingresan candidatos a la arquitectura individualmente si no que se ingresa todo el fotograma para después en la segunda capa convolucional aplicar la capa roi pooling (**Figura 31**) definida en (Girshick, Fast r-cnn, 2015) la cual permite proyectar cada candidato individualmente dentro del mapa de características dado por la última capa convolucional y extraer una región equivalente para luego aplicar max-pool con hiperparámetros variables dependiendo del tamaño del candidato esto con el fin de generar una salida constante tal como se muestra en la **Figura 31**. A diferencia de las arquitecturas basadas en R-CNN, en la cual cada candidato pasa por toda la arquitectura uno cada vez, en esta arquitectura se ingresa a la red convolucional el fotograma y solo en el MLP (el clasificador) se ingresan las características de cada candidato esto para realizar la clasificación final tal como se muestra en la **Figura 30** de la arquitectura propuesta.

En la arquitectura *fast* R-CNN las primeras capas detectan características generales como los bordes tal como se observa en la **Figura 33** la cual representa todos los filtros de la primera capa convolucional, a medida que se aumentan más capas estas detectaran características más específicas como esquinas, cabezas y brazos. La **Figura 34** muestra los primeros 169 filtros de la segunda capa convolucional.

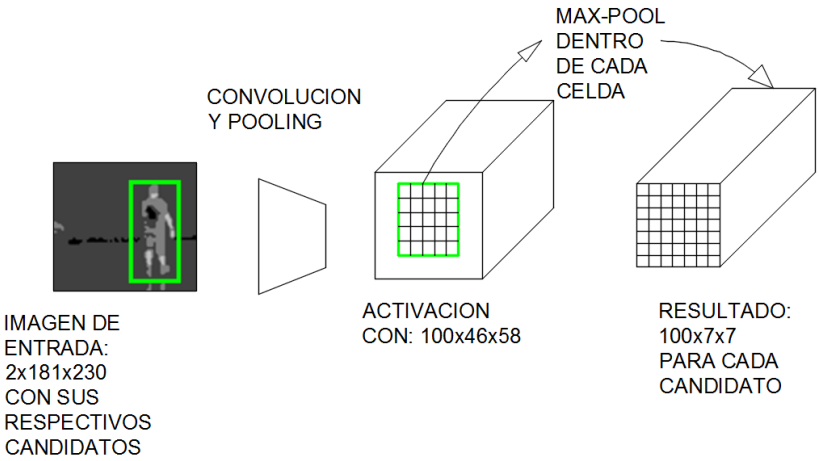


Figura 31. Funcionamiento de la capa ROI pooling.

Fuente: (Girshick, Fast R-CNN, 2015).

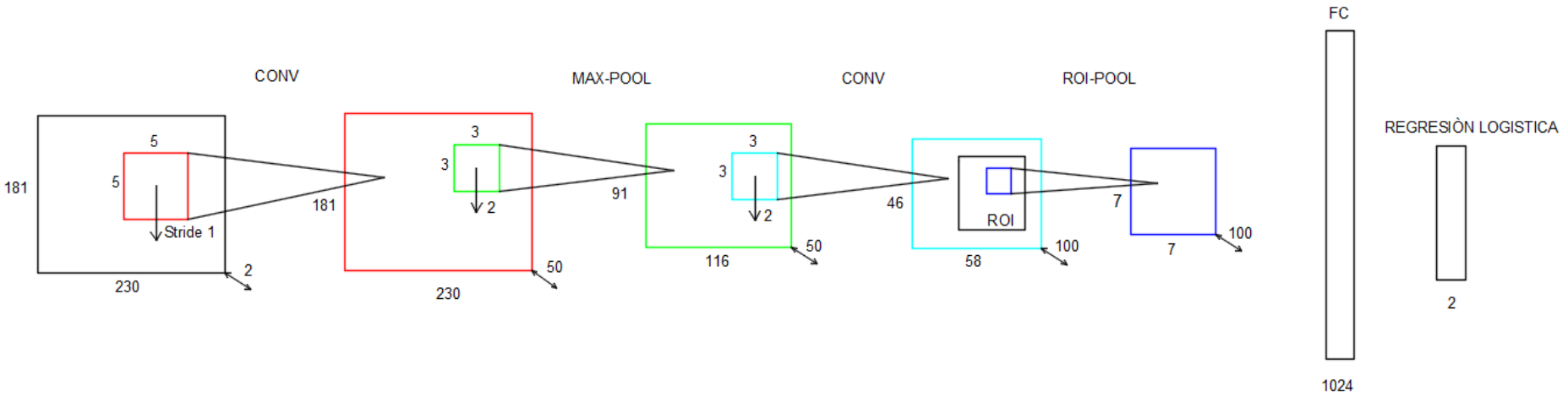


Figura 32. Arquitectura *fast* R-CNN para la detección de peatones por la noche sobre imágenes en el infrarrojo.

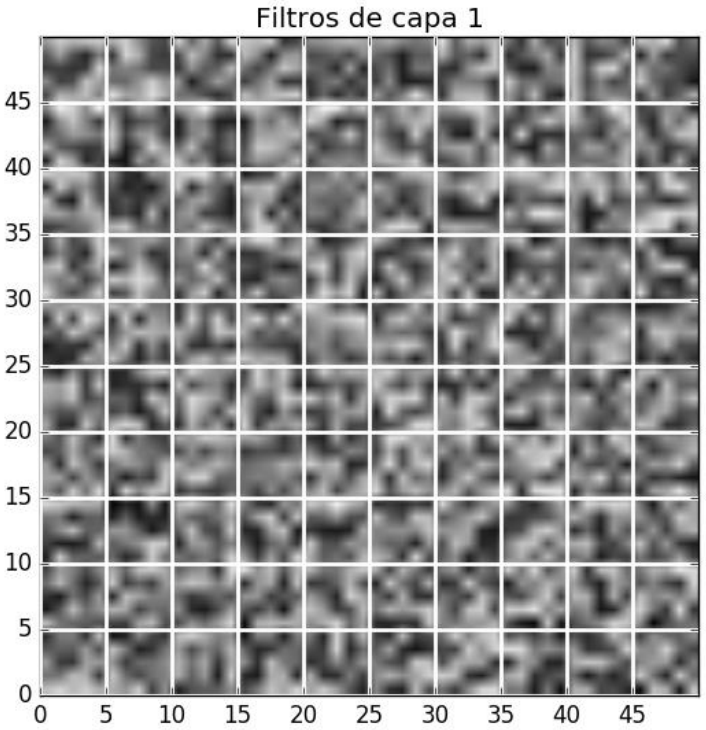


Figura 33. Filtros de la primera capa convolucional.

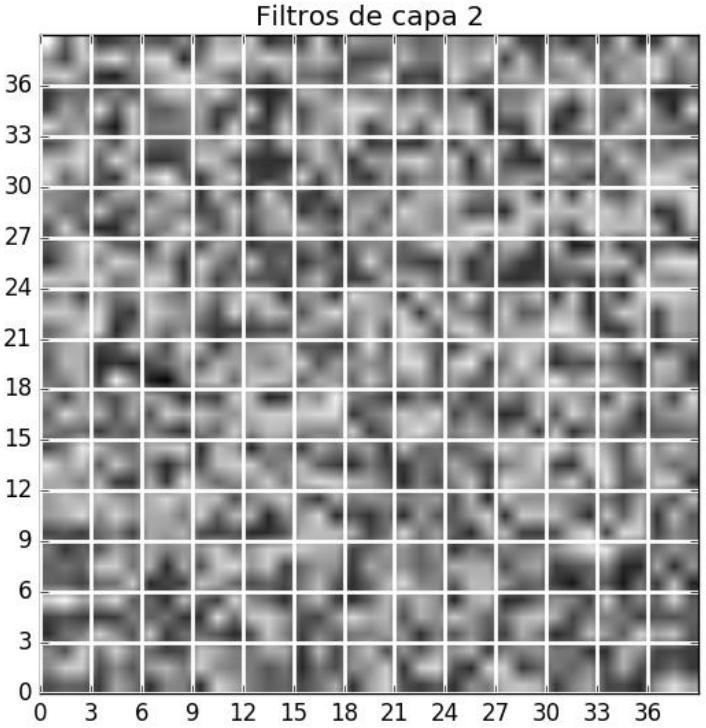


Figura 34. Filtros de la segunda capa convolucional.

4.4 Entrenamiento de la arquitectura *fast* R-CNN para la detección de peatones en la noche

Para este trabajo se utilizaron las bases de datos mencionadas en el estados del arte, de ellas solo se utilizaron imágenes pertenecientes al conjunto Train. Para el entrenamiento se utilizaron candidatos generados con el método descrito en el capítulo anterior. Para reforzar y enriquecer la detección de peatones se utilizaron también imágenes positivas de la base de datos CV-09 FIR de los cuales solo se ocuparon los *ground truth* (ROIS positivos) de estas imágenes.

Tabla 7.
Información de las bases de datos usadas en el entrenamiento de la arquitectura *fast* R-CNN.

	Bases de datos		
	LSI Train	**CV-09 Train**	**Total**
Número de imágenes	3225	2200	5425
Promedio positivos/imagen	3	3	3
Promedio candidatos/imagen	194	3	197
Número de positivos	9675	6600	16275
Número de negativos	615101	0	615101

Del total de datos mostrados en la **Tabla 7.** se utilizaron 20 fotogramas escogidas aleatoriamente en cada época, para cada uno de ellos se utilizaron 4 ejemplos positivos (si es que los había) y 8 ejemplos negativos. Luego de las primeras 50000 épocas se utilizaron 5 fotogramas por cada época, de cada fotograma escogido se utilizaron 4 ejemplos positivos y 100 ejemplos negativos, esto para disminuir la cantidad de falsos positivos que se puedan presentar al probar la arquitectura.

La **Tabla 8** muestra los parámetros que se utilizaron para probar la arquitectura y para determinar las métricas de rendimiento más importantes del detector.

Tabla 8.
Información de las bases de datos usada para determinar métricas de rendimiento.

	LSI Test (clasificación)	**LSI Test (clasificación y detección)**
Número de imágenes	2979	9067
Promedio positivos/imagen	3	3
Promedio candidatos/imagen	371	300
Número de positivos	8937	27201
Número de negativos	1096797	2692899

4.5 Refinación de decisión

En ocasiones se puede dar que se tengan más de una detección por objeto, con lo cual la ventana que mejor se ajuste al objeto y que mayor confidencia tenga se considerará como detección correcta, mientras que las demás detecciones se consideran como falsos positivos lo que merma la calidad del detector. Los métodos de non-maxima supression (NMS) son los encargados de eliminar las múltiples detecciones (redundancias) para un mismo objeto y mantener la detección correcta (Buil, 2011).

En este trabajo se consideran dos algoritmos para afrontar este problema:

- El primero y más conocido se conoce con el nombre de *Greedy* NMS el cual considera las confidencias de las detecciones y el grado de solapamiento entre ellas, este solapamiento se compara con un umbral escogido previamente. Partiendo de una lista de detecciones ordenadas de mayor a menor según su confidencia, compara el solapamiento de cada una de estas

detecciones con las demás y si el solapamiento entre ellas es mayor que un umbral entonces elimina esa detección. Luego iterativamente se realiza el mismo procedimiento con la siguiente detección hasta que no queden elementos en la lista (ver **Figura 35**).

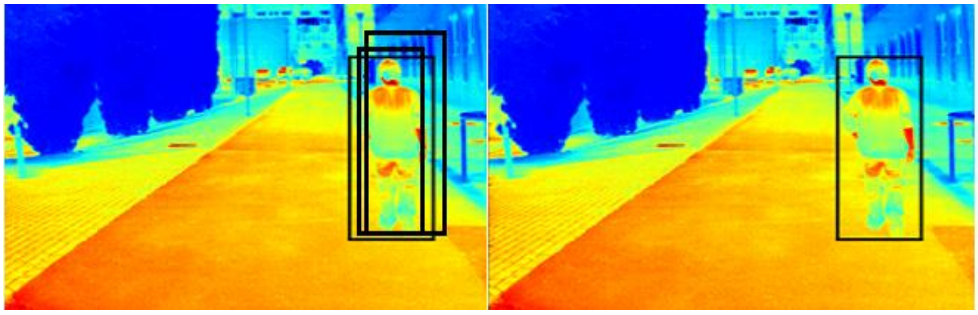


Figura 35. NMS con el método *Greedy* NMS.
Fuente: (Burel & Carel, 1994).

- El segundo algoritmo fue desarrollado en (Felzenszwalb, McAllester, & Ramanan, 2008) y solo considera las detecciones y el grado de solapamiento y no toma en cuenta el grado de certeza. Este algoritmo, según el solapamiento entre dos detecciones, elimina las detecciones que están significativamente cubiertas por una detección previamente seleccionada, tal como se observa en la **Figura 36** lo cual ayuda a eliminar falsos positivos.

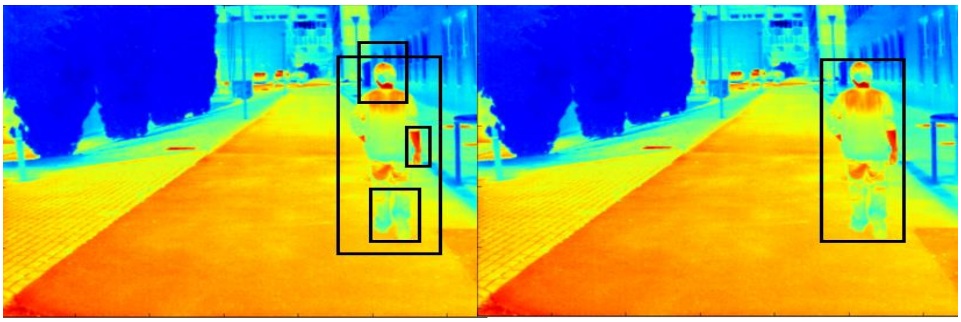


Figura 36. NMS con el método de Felzenszwalb et al. **(Felzenszwalb, McAllester, & Ramanan, 2008)**.
Fuente: (Felzenszwalb, McAllester, & Ramanan, 2008)

4.6 Conclusiones

En este capítulo se ha presentado una nueva arquitectura tipo *fast* R-CNN para el reconocimiento de peatones sobre imágenes en el infrarrojo durante la noche.

Para el entrenamiento se ha utilizado la base de datos LSIFIR Train y CV-09 FIR Train (solo ROIs positivos) para tener una mejor generalización de la arquitectura en la detección de ejemplos positivos.

Se utilizó dos capas convolucionales en la arquitectura al estudiar la arquitectura SPP presentado en (Lee, Ko, & Nam, 2016), se siguió el método de diseño presentado en (Weimer, Scholz-Reiter, & Shpitalni, 2016) el cual determina experimentalmente la arquitectura adecuada para un problema de clasificación.

De los dos métodos de NMS se escogió el segundo sobre el primero porque elimina las posibles falsas detecciones dentro del ROI que define el cuerpo, falsas detecciones como ROIs que enmarquen las extremidades.

CAPÍTULO V
5. PRUEBAS Y RESULTADOS

En este capítulo se presentan los resultados experimentales desarrollados en esta tesis. Los resultados esta divididos en tres partes, en la primera se evalúa el método de generación de ROIs, en la segunda se valora el método de clasificación basado en *fast* R-CNN. Finalmente, en la última etapa se evalúa el sistema detector de peatones de forma global.

5.1 Evaluación del método de generación de ROIs

Para evaluar la etapa de generación de ROIs se siguió los métodos de evaluación descritos en (Kim & Lee, 2013) y (Ge, Luo, & Tei, 2009), donde se evalúa la detección o no detección mediante el solapamiento entre el candidato generado y el *ground truth* (GT) mediante la ecuación (**13)**.

$$solapamiento = \frac{area(deteccion \cap gt)}{area(deteccion \cup gt)} > 0.5 \qquad (13)$$

Esta condición establece que, para que un candidato generado sea considerado como real positivo tiene que existir un solapamiento entre ese candidato y el verdadero, denominado GT de al menos el 50%; si un candidato no tiene correspondencia con ningún GT (la condición expresada en la ecuación (**13)** no se cumple con ningún GT) entonces se considera a ese candidato como falso positivo y si un GT no tiene correspondencia con ningún candidato, entonces la no detección de ese GT se considera un falso negativo.

Con estos parámetros se puede calcular la tasa de error (miss rate) con la ecuación **(14)** y la tasa de detección con la ecuación **(15)** para cada conjunto de valores

que pueden variar principalmente dentro del método de generación de candidatos basado en SURF (Bay, Ess, Tuytelaars, & Van Gool, 2008).

$$tasa\ de\ error = \frac{número\ de\ peatones\ no\ detectados}{número\ total\ de\ peatones\ anotados} \quad (14)$$

$$tasa\ de\ detección = \frac{número\ de\ peatones\ detectados}{número\ total\ de\ peatones\ anotados} \quad (15)$$

El número de candidatos generados por imagen (ROIPI, por sus siglas en inglés Region Of Interest Per Image) depende directamente de los parámetros aplicados al método SURF como el umbral del hessiano, el número de octavas y del número de capas por octava.

Las métricas presentadas en la **Tabla 9** se calcularon sobre el conjunto Test de la base de datos LSIFIR.

Tabla 9.
Rendimiento del método de generación de ROIs.

Umbral hessiano	Octavas	Capas por octava	ROIPI	Tasa de error (%)	Tasa de detección (%)
0.1	9	8	3171.51	0.0995	99.9
0.1	7	6	2999.72	0.4229	99.57
0.2	9	8	2356.2	0.199	99.8
0.2	7	6	2237.9	0.5225	99.47
0.3	9	8	1877.23	0.248	99.75
0.3	7	6	1785.51	0.5722	99.42
0.4	9	8	1503.63	0.3234	99.67
0.4	7	6	1429.04	0.646	99.35
0.5	9	8	1240.85	0.3981	99.60
0.5	7	6	1176.71	0.7962	99.20
0.6	9	8	1049.94	0.47	99.52
0.6	7	6	993.4	0.87	99.12
0.7	9	8	907.82	0.5225	99.47

Continúa

0.7	7	6	857.02	0.9952	99.00
0.8	9	8	801.74	0.6718	99.3
0.8	7	6	755.57	1.1196	98.88
0.9	9	8	710.35	0.920627	99.07
0.9	7	6	668.12	1.368	98.63
1	9	8	638.91	1.19	98.8

En términos generales el objetivo de la generación de ROIs es obtener una tasa de detección alta, con la menor cantidad de ROIPI y con una baja tasa de error. Para escoger un punto de trabajo se tiene que llegar a un compromiso entre la cantidad de ROIPI, la tasa de detección y la tasa de error, ya que cuando se tiene un alto número de ROIPI también se tiene una alta tasa de detección y un valor bajo de tasa de error. En este trabajo se escogió trabajar con (1, 9, 8) como parámetros para el método SURF de generación de ROIs, con lo cual se obtienen 639 candidatos por imagen, 98.8% de tasa de detección con 1.19% de tasa de error.

5.2 Evaluación del método de clasificación

Un clasificador está sujeto a dos tipos de errores, ya sea que este rechace a los ejemplos positivos (falso rechazo o falso negativo) o acepte a los ejemplos negativos (falsa alarma o falso positivo).

En clasificación las métricas más usadas para medir el rendimiento son las curvas ROC (por sus siglas en inglés Receiver Operating Characteristic) (Spackman, 1989) y DET (por sus siglas en inglés Detection Error Tradeoff) (Martin, Doddington, Kamm, Ordowski, & Przybocki, 1997).

Las curvas ROC ayudan a determinar qué tan buena es la clasificación, está basada en el número de clasificaciones correctas.

Las curva ROC es una gráfica que permite evaluar el rendimiento de problemas de clasificación en los cuales existe una variable que puede ser admitida dentro de dos categorías, y presenta en una gráfica la tasa de reales positivos (true positive rate)

versus la tasa de falsos positivos (false positive rate) para todos los diferentes posibles umbrales de clasificación. Con esta gráfica se pueden evaluar diferentes clasificadores según el área que cada uno tenga bajo la curva, mientras el área este más cercano a uno, su rendimiento va a ser mejor. Esta gráfica ayuda también a determinar cuál es el umbral que mejores resultados presenta, se escoge el valor de umbral que dé como resultado el punto más cercano a uno dentro de la tasa de reales positivos.

Dado una matriz de confusión (Pearson, 1904) como la **Tabla 10**, la tasa de reales positivos y la tasa de falsos positivos se calculan con las siguientes ecuaciones:

$$sensibilidad\ o\ tasa\ de\ reales\ positivos = \frac{RP}{RP + FN} \tag{16}$$

$$tasa\ de\ falsos\ positivos = \frac{FP}{RN + FP} \tag{17}$$

$$tasa\ de\ falsos\ negativos = 1 - sensibilidad \tag{18}$$

Tabla 10.
Matriz de confusión para clasificación de peatones.

		Resultado clasificación	
		Peatón	No peatón
Instancia real	Peatón	Reales positivos (RP)	Falsos negativos (FN)
	No peatón	Falsos positivos (FP)	Reales negativos(RN)

En este trabajo las características generadas se clasifican con regresión logística, la regresión logística se utiliza para predecir la probabilidad de que alguna observación pertenezca a la clase "peatón", dado un vector de características $X \in R^{1024}$ es decir se obtiene la probabilidad condicional dadas por las ecuaciones **(19)** y **(20).** Para tener una visualización de los ejemplos clasificados, mediante PCA se redujo la dimensionalidad del vector de características a R^2 y se graficó los ejemplos categorizados en la **Figura 37**.

$$P(y = peaton|\, X) = \frac{1}{1 + e^{-\theta^T X}} \tag{19}$$

$$P(y = no\ peaton|\, X) = 1 - P(y = peaton|X) \tag{20}$$

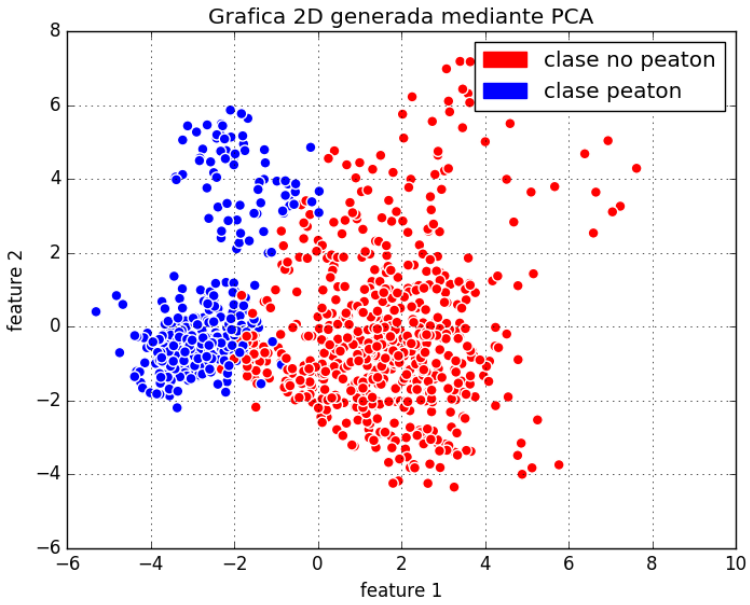


Figura 37. Categorización de los ejemplos de la base de datos LSI FIR.

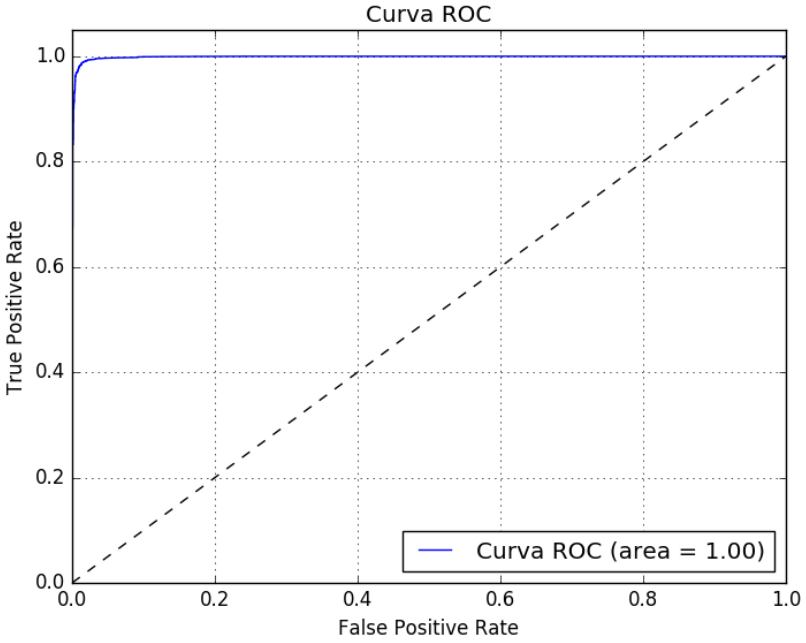


Figura 38. Curva ROC para el clasificador de peatones en la noche basado en *fast* R-CNN.

Para evaluar la clasificación una métrica comúnmente utilizada es el área bajo la curva (AUC, por sus siglas en inglés Area Under the Curve) el cual indica que tan buena es la clasificación de ejemplos positivos en función de la falsa aceptación de ejemplos negativos, este AUC es aproximadamente 1 para este trabajo.

Las curvas DET determinan que tan bueno es el rechazo de ejemplos negativos y está basado en número de rechazos correctos.

La curva DET en clasificación presenta en una gráfica logarítmica la tasa de falsos negativos (false negative rate) versus la tasa de falsos positivos (false positive rate) para todos los diferentes posibles umbrales.

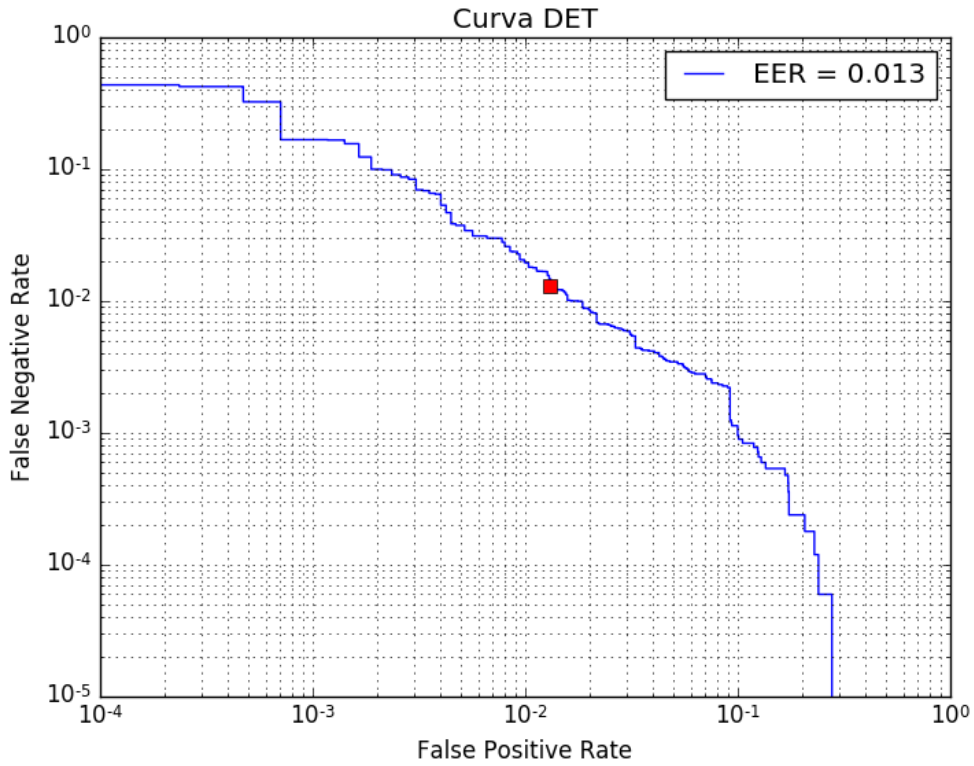


Figura 39. Curva DET para el clasificador de peatones en la noche basado en *fast* R-CNN sobre la base de datos LSIFIR.

Existen dos formas de interpretar la curva DET, la primera mediante el área bajo la curva y la segunda y más comúnmente usada mediante el EER (por sus siglas en inglés Equal Error Rate) aunque otros trabajos como (Olmeda, Premebida, Nunes,

Armingol, & de la Escalera, 2013) utilizan una tercera forma de evaluación la cual consiste en evaluar la curva en 10^{-4} (si es que existe este punto) y obtener la tasa de falsos negativos en ese punto. El EER es el punto donde la tasa de falsos positivos (tasa de aceptación) y la tasa de falsos negativos (tasa de rechazo) son iguales, para este trabajo es de aproximadamente 1.3% (mientras menor sea este número mejor es el rendimiento).

Estas dos gráficas se calcularon con 1000 fotogramas positivos escogidos al azar utilizando los datos mostrados en la **Tabla 8** de la parte de clasificación, de los cuales se escogió 5 ejemplos negativos y todos los ejemplos positivos disponibles por cada fotograma.

5.3 Evaluación del método de detección de peatones en la noche sobre imágenes en el infrarrojo

Para evaluar el rendimiento global del detector se utilizó un esquema de evaluación por fotograma utilizando una versión modificada del esquema presentado en los retos de detección de objetos de PASCAL (Everingham, Van, Williams, Winn, & Zisserman, 2010) utilizada en (Dollar, Wojek, Schiele, & Perona, 2012), específicamente se calcularon dos gráficas. En la primera grafica se representó la tasa de detección versus el número de falsos positivos por imagen (FPPI), y en la segunda se tiene la tasa de error versus el número de falsos positivos por imagen, todos estos valores se calcularon con valores de umbrales entre cero y uno. Esta evaluación se realizó sobre la lista de ROIs generada en la salida de la etapa NMS la cual une detecciones similares. Todos estos parámetros se calcularon con las ecuaciones (**14)**, (**21)** y (**22)**.

$$tasa\ de\ detección = \frac{número\ de\ peatones\ clasificados\ correctamente}{número\ total\ de\ peatones\ anotados} \tag{21}$$

$$fppi = \frac{número\ de\ falsos\ positivos}{número\ de\ fotogramas\ usados} \tag{22}$$

Para este trabajo se consideró que una detección es correcta cuando el solapamiento entre este y algún GT (los peatones anotados) es mayor a 0.5, el solapamiento se define según la ecuación (**13)**. El solapamiento representa la relación entre el área de la intersección entre la detección y el GT y el área de la unión entre la detección y el mismo GT. Específicamente para calcular las dos graficas mencionadas se siguió el algoritmo descrito en (Russakovsky, 2015) donde se define el real positivo, falso positivo y falso negativo para detección según la ecuación (**23**).

$$\begin{gathered} RP\ si\ solapamiento(Candidato, GT) > 0.5 \\ FP\ si\ solapamiento(Candidato, GT) \leq 0.5 \\ FN\ si\ GT\ no\ tiene\ ningun\ candidato\ asociado \end{gathered} \tag{23}$$

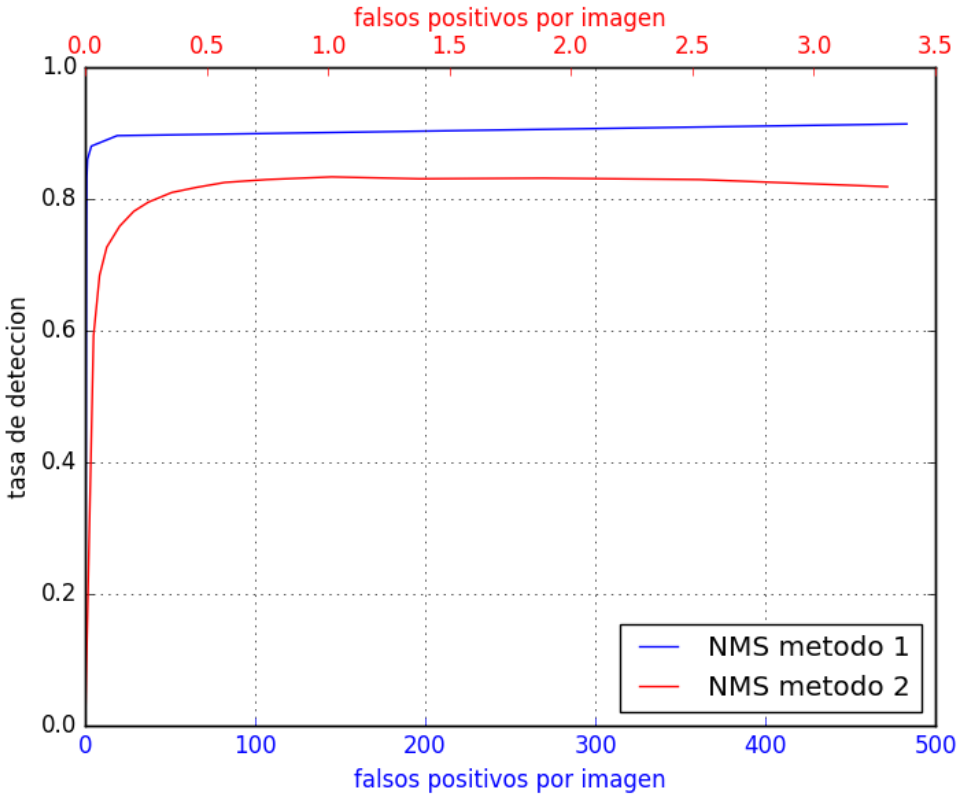


Figura 40. Curva FPPI vs tasa de detección para la detección de peatones en la noche sobre imágenes en el infrarrojo utilizando los dos métodos de *non-maxima supression* (NMS) descritos en el capítulo anterior.

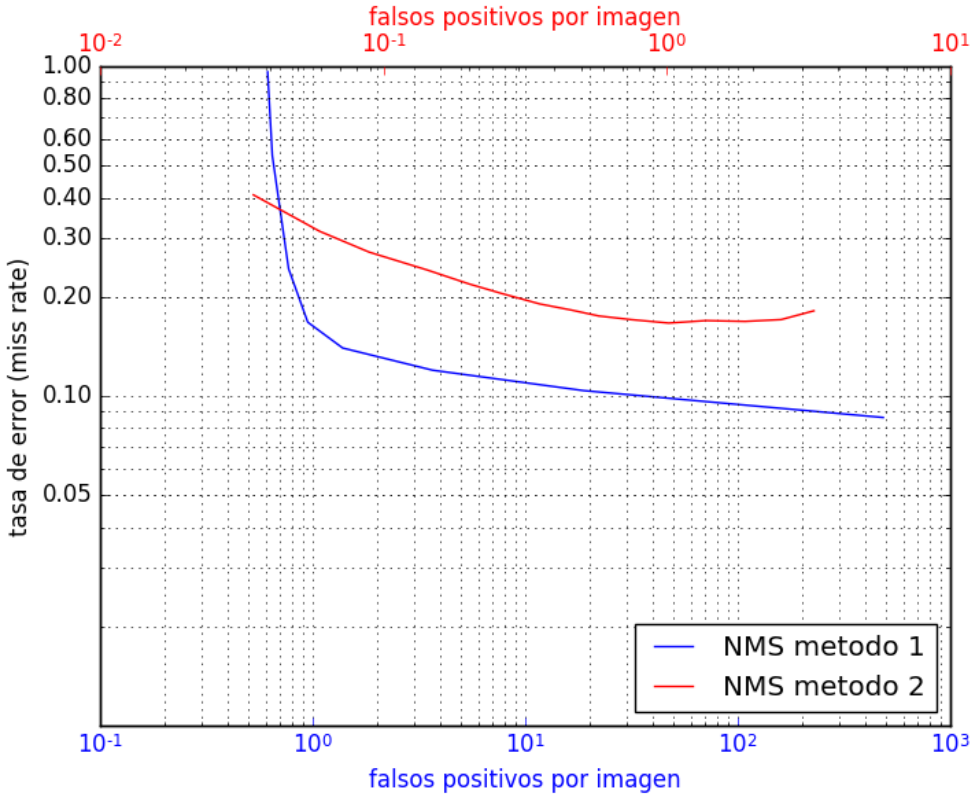


Figura 41. Curva DET para para la detección de peatones en la noche sobre imágenes en el infrarrojo utilizando los dos métodos de *non-maxima supression* (NMS) descritos en el capítulo anterior.

Tabla 11.
Comparación de la tasa de error en la detección de peatones en la noche a 10^{-1}.

	Tasa de error
Método propuesto	0.255
(Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013)	0.251
(John, Mita, Liu, & Qi, 2015)	0.654

Para resumir el rendimiento del detector y para comparar con otros sistemas se utilizó el método descrito en (Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013) el cual evalúa la tasa de error en 10^{-1} FPPI. Se comparó el sistema usando los dos métodos NMS mencionados en el capítulos anterior, se deduce que con el segundo método de NMS se obtiene una menor cantidad de falsos positivos sacrificando muy poca tasa de detección.

Como era de esperarse la tasa de detección disminuye en comparación con la curva ROC, debido a que esta curva toma en cuenta el método de generación de regiones de interés. En (Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013) se calcula la tasa de error que engloba el rendimiento global del sistema para 10^{-1} dando como resultado 0.251 un rendimiento que se compara con el obtenido de 0.255 en el método propuesto, mientras que en (John, Mita, Liu, & Qi, 2015) utilizando esta misma base de datos se obtiene 0.657 como tasa de error. Esta métrica se calculó sobre toda la base de datos Test de LSI FIR.

5.4 Desempeño del sistema de detección

Se escogió un punto de trabajo con umbral de clasificación de 0.5 y se generó los resultados mostrados en la Tabla 12.

Tabla 12.
Rendimiento del sistema de detección de peatones en la noche.

Nombre	Número de imágenes	FPPI	Tasa de detección (%)	Miss rate (%)
LSI Test	9067	0.7	83	17

5.5 Tiempo de procesamiento

Los algoritmos fueron implementados en C++, usando las librerías Caffe-Fast y OpenCV 3.1.

Para el desarrollo de los experimentos se utilizó un computador de escritorio Intel Core i7-4790 CPU @ 3.60GHz y 8 GB de RAM con sistema operativo Ubuntu 14.04 de 64 bits.

El tiempo de procesamiento es de aproximadamente 3 fotogramas por segundo trabajando con fotogramas de tamaño 181 x 230 píxeles. Este tiempo de computación depende del número de candidatos que se generen en cada fotograma y podría disminuir utilizando alguna técnica de paralización.

5.6 Conclusiones

La evaluación del método de generación de ROIs indica que son muy pocos los candidatos que se pasan por alto, es por esto que se concluye que el rendimiento del sistema depende principalmente del método de reconocimiento o clasificación.
El segundo método de NMS elimina una gran cantidad de falsos positivos sacrificando muy poca tasa de detección llegando a obtener una tasa de error de 25.5% comparable con (Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013) y que supera la tasa de error obtenida en el trabajo de (John, Mita, Liu, & Qi, 2015).

CAPÍTULO VI
6. CONCLUSIONES Y RECOMENDACIONES

En este capítulo final se presentan las conclusiones y las recomendaciones a las que se han llegado al final de esta investigación.

6.1 Conclusiones

Este trabajo presenta una nueva arquitectura *fast* R-CNN con dos capas convolucionales para la detección de peatones en la noche usando imágenes del infrarrojo lejano. Para diferenciar un peatón de un no peatón la arquitectura genera un conjunto de características sobre todo el fotograma, las mismas que se calculan una sola vez, por lo que para clasificar cada candidato simplemente se extrae del mapa de características generado en la última capa convolucional la región correspondiente al candidato y se lo categoriza a través de regresión logística. Esta arquitectura muestra que con apenas dos capas convolucionales se alcanza una tasa de error de 25.5% comparable con el método que genera características manuales presentado en (Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013) el cual alcanza una tasa de error de 25.1%.

Los métodos basados en CNN requieren la aplicación de varias capas de procesamiento aplicadas jerárquicamente por lo cual en general requieren más tiempo de procesamiento que los métodos que generan características manuales por lo que para que se ejecuten en tiempo real requieren de una implementación o codificación especializada que optimice los recursos que se tengan disponibles como los núcleos de un CPU combinado con el uso de tarjetas gráficas (GPU).

La principal ventaja de esta arquitectura radica en que las características de todos los posibles candidatos ya se encuentran calculadas en la última capa convolucional a diferencia de las arquitecturas basadas en R-CNN en las cuales se extraen características por candidato individualmente, es decir se hace pasar a cada candidato

por toda la arquitectura uno cada vez. Otra ventaja es que la arquitectura utilizada elimina la necesidad de fijar la resolución de los candidatos a priori, esto gracias a la capa roi *pooling*.

En la generación de candidatos se utilizó un detector de puntos de interés que permite obtener los cuerpos claros de la imagen, estos cuerpos claros tienen alto probabilidad de ser la cabeza de un peatón, según esto se calcula la región de interés en base al tamaño del cuerpo, en la generación de candidatos con alto contraste con el fotograma solo se tomaron en cuenta aquello candidatos con un área mayor a 150 pixeles y con una relación de aspecto entre 0.2 y 0.75. En este trabajo se escogió (1, 9, 8) como parámetros para el método SURF de generación de ROIs con bajo contraste con el fotograma, con lo cual se obtiene 639 candidatos por imagen aproximadamente, 98.8% de tasa de detección de ROIs con 1.44% de tasa de error de ROIs sobre la base de datos LSIFIR.

La evaluación del método de generación de ROIs presenta una tasa de detección del 98.8% por lo cual el rendimiento del detector depende principalmente de la etapa de reconocimiento o clasificación. En general según las gráficas presentadas en el capítulo anterior, no existe mayor problema en la detección de ejemplos positivos siempre y cuando la detección exista, esto se puede apreciar en la curva FPPI versus tasa de detección que toma en cuenta la detección del candidato por el método de generación de regiones de interés.

Los principales problemas que se pueden apreciar en la gráfica son la gran cantidad de FPPI usando el primer método de NMS, lo cual invita a escoger un umbral de clasificación alto que permita filtrar esos falsos positivos, el otro problema es la no detección (tasa de error) el cual depende directamente del método de clasificación. El problema de la gran cantidad de FPPI se corrige con el segundo método de NMS.

En los sistemas ADAS de detección de peatones generalmente se tiene un número máximo de FPPI admisible en el sistema es por esto que se utiliza la curva tasa de error versus FPPI para comparar sistemas y para determinar un punto de trabajo.

En la **Tabla 11** se contrasta este trabajo con otros sistemas, donde se muestra una tasa de error que se compara con el obtenido en (Olmeda, Premebida, Nunes, Armingol, & de la Escalera, 2013) y que supera el de (Socarras, Ramos, Vazquez, Lopez, & Gevers, 2011) utilizando una arquitectura pequeña de dos capas convolucionales comparada con las arquitecturas CNN mencionadas en el estado del arte donde la más pequeña tiene al menos cinco capas convolucionales, por estas razones se concluye que generar características puntuales mediante CNN es más adecuado que generar características con métodos generales manualmente y clasificarlas mediante SVM o cualquier otro clasificador el cual ha sido el enfoque predominante en los últimos años.

El tiempo de procesamiento depende de la cantidad de candidatos que se generan en cada fotograma, si se utilizan los dos métodos de generación de ROIs se procesan 3 fotogramas por segundo, conociendo previamente el ambiente donde se va a emplear el sistema, se podría limitar la cantidad de candidatos generados utilizando solamente el método de regiones crecientes y se puede llegar a procesar hasta 13 fotogramas por segundo.

6.2 Recomendaciones

Durante las pruebas del sistema se determinaron acciones que pueden mejorar este trabajo:

- Aumentar la complejidad de la arquitectura en el caso que se tenga el hardware necesario (una tarjeta gráfica NVIDIA compatible con CUDA) que soporte aquello.
- Trabajar con fotogramas de mayor tamaño con el fin de detectar candidatos más pequeños en largas distancias, especialmente, para aplicaciones en detección de peatones en la noche, donde es importante la detección con suficiente antelación.
- Incorporar un módulo de seguimiento basado en el Filtro Kalman, para resolver dos problemas, reducir el tiempo de procesamiento y la estimar la trayectoria del peatón.

- La mayor parte del tiempo de procesamiento se genera en la etapa de clasificación por lo que se podría aprovechar la programación paralela implementada en la librería (llamada caffe) de aprendizaje automático utilizada, para disminuir el tiempo de procesamiento lo cual requiere la utilización de una tarjeta gráfica NVIDIA compatible con CUDA.

REFERENCIAS BIBLIOGRÁFICAS

¿Qué son los sistemas de visión nocturna? (9 de Mayo de 2013). Recuperado el 19 de Noviembre de 2016, de http://www.circulaseguro.com/que-son-los-sistemas-de-vision-nocturna/

Agirregabiria, M. (Febrero de 2012). *Curva sigmoidea*. Recuperado el 20 de Noviembre de 2016, de http://blog.agirregabiria.net/2012/02/curva-sigmoidea-el-secreto-de-la-vida-y.html

Alpaydin, E. (2009). *Introduction to Machine Learning.* MIT press.

An Intuitive Explanation of Convolutional Neural Networks. (11 de Agosto de 2016). Recuperado el 15 de Febrero de 2017, de https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

ANT. (2016). *Estadisticas de Transporte Terrestre y Seguridad Vial.* Recuperado el 17 de Noviembre de 2016, de http://www.ant.gob.ec/index.php/noticias/estadisticas#.WDDoINFVK1E

ANT. (2016). *Siniestros octubre 2016.* Recuperado el 14 de Noviembre de 2016, de http://www.ant.gob.ec/index.php/descargable/file/3813-siniestros-octubre-2016

(2014). *Anuario de Estadísticas de Transportes 2013.* Obtenido de http://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Economicas/Estadistica%20de%20Transporte/Publicaciones/Anuario_de_Estad_de_Transporte_2013.pdf

(2015). *Anuario de Estadísticas y Transportes 2014.* Obtenido de http://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Economicas/Estadistica%20de%20Transporte/Publicaciones/Anuario_de_Estad_de_Transporte_2014.pdf

Arel, I., Rose, D., & Karnowski, T. P. (2010). Deep machine learning-a new frontier in artificial intelligence research. *IEEE Computational Intelligence Magazine, 5*(4), 13-18.

Artificial intelligence in process automation. (2006). Obtenido de http://www.controlglobal.com/articles/2006/221/

Baker, B., Gupta, O., Naik, N., & Raskar, R. (2016). Designing Neural Network Architectures using Reinforcement Learning. *arXiv preprint arXiv:1611.02167*.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer vision and image understanding, 110*(3), 346-359.

Besbes, B., Rogozan, A., Rus, A.-M., Bensrhair, A., & Broggi, A. (2015). Pedestrian detection in far-infrared daytime images using a hierarchical codebook of SURF. *Sensors, 15*(4), 8570--8594.

Brief introduction to Intelligent Transportation System, ITS. (s.f.). Recuperado el 15 de Noviembre de 2016, de https://www.freeway.gov.tw/UserFiles/File/Traffic/A1%20Brief%20introduction%20to%20Intelligent%20Transportation%20System,%20ITS.pdf

Buil, M. D. (2011). *Non-Maxima Supression.* Graz: Inst. for Computer Graphics and Vision.

Burel, G., & Carel, D. (1994). Detection and localization of faces on digital images. *Pattern Recognition Letters, 15*(10), 963--967.

Chen, M. (2015). *Pedestrian detection with RCNN.*

Cisneros, Ó. (2008). *e-Safety: Nuevas tecnologías al servicio de la seguridad vial*. Recuperado el 18 de Noviembre de 2016, de El sistema de visión nocturna: http://www.centro-zaragoza.com:8080/web/sala_prensa/revista_tecnica/hemeroteca/articulos/R37_A7.pdf

Criminisi, A., & Shotton, J. (2013). *Decision forests for computer vision and medical image analysis.*

Curio, C., Edelbrunner, J., Kalinke, T., Tzomakas, C., & Von Seelen, W. (2000). Walking pedestrian recognition. (IEEE, Ed.) *IEEE Transactions on intelligent transportation systems, 1*(3), 155-163.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1*, 886-893.

Davis, J. W., & Keck, M. A. (2005). A Two-Stage Template Approach to Person Detection in Thermal Imagery. *WACV/MOTION, 5*, 364-369.

Delgado Landázuri, L. F. (2015). Diseño y construcción de una plataforma inteligente y portátil para un sistema de visión estéreo para aplicación de visión por computadora en vehículos. ESPE.

Deng, L. (2014). A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing, 3*(e2).

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence, 34*(4), 743-761.

Dong, J., Ge, J., & Luo, Y. (Septiembre de 2007). Nighttime pedestrian detection with near infrared using cascaded classifiers. *2007 IEEE International Conference on Image Processing*, *6*, VI-185-VI-188.

Duchon, F., Hubinský, P., Babinec, A., & Tölgyessy, M. (2012). (Elsevier, Ed.) *Intelligent vehicles as the robotic applications*, *48*, 105-114.

Electromagnetic Radiation. (2001). Recuperado el 15 de Noviembre de 2016, de Electromagnetic Waves: http://www.crisp.nus.edu.sg/~research/tutorial/em.htm

Enzweiler, M., & Gavrila, D. M. (2011). A multilevel mixture-of-experts framework for pedestrian classification. (IEEE, Ed.) *IEEE Transactions on Image Processing*, *20*(10), 2967-2979.

Everingham, M., Van, G. L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, *88*(2), 303--338.

Everingham, M., Van-Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (18 de Marzo de 2017). *The PASCAL Visual Object Classes Challenge 2007*. Obtenido de http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

Experimento de Herschel en la banda infrarroja. (s.f.). Recuperado el 16 de Noviembre de 2016, de http://legacy.spitzer.caltech.edu/espanol/edu/herschel/experiment.shtml

(s.f.). *Fallecidos Octubre 2016.* Recuperado el 17 de Noviembre de 2016, de http://www.ant.gob.ec/index.php/descargable/file/3812-fallecidos-octubre-2016

Fang, Y., Yamada, K., Ninomiya, Y., Horn, B. K., & Masaki, I. (2004). A shape-independent method for pedestrian detection with far-infrared images. (IEEE, Ed.) *IEEE Transactions on Vehicular Technology*, *53*(6), 1679-1697.

Fang, Y., Yamada, K., Ninomiya, Y., Horn, B., & Masaki, I. (2003). Comparison between infrared-image-based and visible-image-based approaches for pedestrian detection. *Intelligent Vehicles Symposium*, 505-510.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1--8.

Ge, J., Luo, Y., & Tei, G. (2009). Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *IEEE Transactions on Intelligent Transportation Systems*, *10*(2), 283-298.

Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 1440-1448.

Girshick, R. (13 de Octubre de 2015). *Fast R-CNN*. Obtenido de http://web.cs.ucdavis.edu/~yjlee/teaching/ecs289g-fall2015/charlie1.pdf

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580-587.

Github. (3 de Marzo de 2017). Obtenido de https://github.com/rbgirshick/voc-dpm/blob/master/test/nms.m

Goodfellow, I., Bengio, Y., & Courville, A. (2015). *Deep learning.*

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27-48.

Haar-feature Object Detection in C#. (8 de Diciembre de 2016). Obtenido de https://www.codeproject.com/Articles/441226/Haar-feature-Object-Detection-in-Csharp

Hammond, M., Qu, G., & Rawashdeh, O. A. (Agosto de 2015). Deploying and Scheduling Vision Based Advanced Driver Assistance Systems (ADAS) on Heterogeneous Multicore Embedded Platform. *2015 Ninth International Conference on Frontier of Computer Science and Technology*, 172-177.

He, K., Zhang, X., Ren, S., & Sun, J. (Septiembre de 2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. *European Conference on Computer Vision*, 346-361.

INEC. (2015). *Anuario de estadísticas vitales nacimientos y defunciones 2014.* Recuperado el 14 de Noviembre de 2016, de Ecuador en cifras: http://www.ecuadorencifras.gob.ec/documentos/web-inec/Poblacion_y_Demografia/Nacimientos_Defunciones/Publicaciones/Anuario_Nacimientos_y_Defunciones_2014.pdf

Infrared. (16 de Abril de 2014). Recuperado el 18 de Noviembre de 2016, de http://www.newworldencyclopedia.org/entry/Infrared

Infrared Explained. (s.f.). Recuperado el 20 de Noviembre de 2016, de http://floor-heating.co.uk/infrared-explained/

Introducción a las redes neuronales. (s.f.). Recuperado el 20 de Noviembre de 2016, de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema3dm.pdf

Jain, A. (4 de Abril de 2016). *Deep Learning for Computer Vision – Introduction to Convolution Neural Networks*. Obtenido de https://www.analyticsvidhya.com/blog/2016/04/deep-learning-computer-vision-introduction-convolution-neural-networks/

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., . . . Darrell, T. (Noviembre de 2014). Caffe: Convolutional architecture for fast feature

embedding. *Proceedings of the 22nd ACM international conference on Multimedia*, 675-678.

John, V., Mita, S., Liu, Z., & Qi, B. (2015). Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks. *Machine Vision Applications (MVA), 2015 14th IAPR International Conference*, 246-249.

Karpathy, A. (s.f.). *Convolutional Neural Networks*. Recuperado el 21 de Noviembre de 2016, de http://cs231n.github.io/

Kharchenko, V., Orehov, A., Brezhnev, E., Orehova, A., & Manulik, V. (Septiembre de 2014). The cooperative human-machine interfaces for cloud-based advanced driver assistance systems: Dynamic analysis and assurance of vehicle safety. *Design & Test Symposium (EWDTS), 2014 East-West*, 1-5.

Kim, D. S., & Lee, K. H. (2013). Segment-based region of interest generation for pedestrian detection in far-infrared images. (Elsevier, Ed.) *Infrared Physics & Technology, 61*, 120-128.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097-1105.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

Lee, E. J., Ko, B. C., & Nam, J. Y. (2016). Recognizing pedestrian's unsafe behaviors in far-infrared imagery at night. *Infrared Physics & Technology, 76*, 261-270.

Li, J., Liang, X., Shen, S., Xu, T., Feng, J., & Yan, S. (2015). Scale-aware fast R-CNN for pedestrian detection. *arXiv preprint arXiv:1510.08160*.

Li, L., & Zhu, X. (Enero de 2013). Design concept and method of advanced driver assistance systems. *2013 Fifth International Conference on Measuring Technology and Mechatronics Automation*, 434-437.

Lim, J. H., Tsimhoni, O., & Liu, Y. (2010). Investigation of driver performance with night vision and pedestrian detection systems—Part I: Empirical study on visual clutter and glance behavior. *IEEE Transactions on Intelligent Transportation Systems, 11*(3), 670-677.

Lin, C.-F., Chen, C.-S., Hwang, W.-J., Chen, C.-Y., Hwang, C.-H., & Chang, C.-L. (2015). Novel outline features for pedestrian detection system with thermal images. *Pattern Recognition, 48*(11), 3440-3450.

Liu, Q., Zhuang, J., & Ma, J. (2013). Robust and fast pedestrian detection method for far-infrared automotive driving assistance systems. (Elsevier, Ed.) *Infrared Physics & Technology, 60*, 288-299.

Los accidentes de tránsito son un problema socioeconomico. (s.f.). Recuperado el 12 de Noviembre de 2016, de http://sanpedrodelapaz.cl/wp-content/uploads/2013/10/accidentes_transito.pdf

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision, 60*(2), 91-110.

LSI Far Infrared Pedestrian Dataset. (9 de Diciembre de 2016). Obtenido de http://portal.uc3m.es/portal/page/portal/dpto_ing_sistemas_automatica/investigacion/IntelligentSystemsLab/research/InfraredDataset

Lucas, J. (26 de Marzo de 2015). *What Is Infrared?* Recuperado el 17 de Noviembre de 2016, de http://www.livescience.com/50260-infrared-radiation.html

Ma, Y., Chen, X., & Chen, G. (2011). Pedestrian detection and tracking using HOG and oriented-LBP features. *IFIP International Conference on Network and Parallel Computing*, 176-184.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The det curve in assessment of detection task performance. *Proc. Eurospeech 97*, 1895-1898.

Ng, A. (s.f.). *CS229 Lecture notes.* Obtenido de http://cs229.stanford.edu/notes/cs229-notes1.pdf

Ng, A., Ngiam, J., Foo, C. Y., Mai, Y., Suen, C., Coates, A., . . . Tandon, S. (s.f.). *Deep Learning Tutorial*. Recuperado el 20 de Noviembre de 2016, de http://ufldl.stanford.edu/tutorial/

O'Malley, R., Jones, E., & Glavin, M. (2010). Detection of pedestrians in far-infrared automotive night vision using region-growing and clothing distortion compensation. *Infrared Physics & Technology, 53*(6), 439-449.

Olmeda, D. (2014). *Pedestrian Detection in Infrared Images.* Leganés.

Olmeda, D., de la Escalera, A., & Armingol, J. M. (2012). Contrast invariant features for human detection in far infrared images. *Intelligent Vehicles Symposium (IV), 2012 IEEE*, 117-122.

Olmeda, D., Premebida, C., Nunes, U., Armingol, J. M., & de la Escalera, A. (2013). Pedestrian detection in far infrared images. *Integrated Computer-Aided Engineering*, 347-360.

OMS. (2013). *Seguridad peatonal*. Recuperado el 10 de Mayo de 2017, de Manual de seguridad vial para instancias decisorias y profesionales: http://apps.who.int/iris/bitstream/10665/128043/1/9789243505350_spa.pdf?ua=1

OMS. (2015). *Informe sobre la situación mundial de la seguridad vial 2015.* Recuperado el 24 de Noviembre de 2016, de

http://www.who.int/violence_injury_prevention/road_safety_status/2015/Summary_GSRRS2015_SPA.pdf?ua=1

OMS. (s.f.). *Lesiones causadas por el tránsito.* Recuperado el 14 de Noviembre de 2016, de http://www.who.int/mediacentre/factsheets/fs358/es/

Pearson, K. (1904). *Mathematical Contributions to the Theory of Evolution.*

Proporciones del cuerpo humano. (27 de Abril de 2013). Recuperado el 12 de Diciembre de 2016, de http://www.comocubriruncuerpo.org/proporciones-del-cuerpo-humano-1-da-vinci-y-vitruvio/

Qi, B., John, V., Liu, Z., & Mita, S. (2016). Pedestrian detection from thermal images: A sparse representation based approach. *Infrared Physics & Technology, 76*, 157-167.

Redes neuronales artificiales. (s.f.). Recuperado el 19 de Noviembre de 2016, de https://www.emaze.com/@ALFOQZWI/Redes-Neuronales-Artificiales

Robles-Kelly, A., & Huynh, C. P. (2013). *Imaging Spectroscopy for Scene Analysis.* Springer.

Russakovsky, O. (2015). *Scaling up object detection.* Obtenido de http://ai.stanford.edu/~olga/papers/PhD_thesis.pdf

Schalkoff, R. J. (1997). *Artificial Neural Networks.* McGraw-Hill.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning.* Cambridge University Press.

Shaout, A., Colella, D., & Awad, S. (Diciembre de 2011). Advanced driver assistance systems-past, present and future. *Computer Engineering Conference (ICENCO), 2011 Seventh International*, 72-82.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Socarras, Y., Ramos, S., Vazquez, D., Lopez, A. M., & Gevers, T. (Diciembre de 2011). Adapting Pedestrian Detection from Synthetic to Far Infrared Images. *Proceedings of the International Conference on Computer Vision, Workshop on Visual Domain Adaptation and Dataset Bias, Sydney, Australia, 7*.

Soga, M., Hiratsuka, S., Fukamachi, H., & Ninomiya, Y. (2008). Pedestrian detection for a near infrared imaging system. *2008 11th International IEEE Conference on Intelligent Transportation Systems*, 1167-1172.

Spackman, K. A. (1989). Signal detection theory: Valuable tools for evaluating inductive learning. *Proceedings of the sixth international workshop on Machine learning*, 160-163.

Suard, F., & Rakotomamonjy, A. (2006). Pedestrian detection using infrared images and histograms of oriented gradients. *Intelligent Vehicles Symposium, 2006 IEEE*, 206-212.

Szarvas, M., Sakai, U., & Ogata, J. (2006). Real-time pedestrian detection using LIDAR and convolutional neural networks. *2006 IEEE Intelligent Vehicles Symposium*, 213-218.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*, 1-9.

Theano. (2013). *Convolutional Neural Networks (LeNet)*. Recuperado el 20 de Marzo de 2017, de http://deeplearning.net/tutorial/lenet.html

Tome, D., Monti, F., Baroffio, L., Bondi, L., Tagliasacchi, M., & Tubaro, S. (2016). Deep convolutional neural networks for pedestrian detection. *Signal Processing: Image Communication.*

Vanajakshi, L., Ramadurai, G., & Anand, A. (2010). *Intelligent transportation system.* Obtenido de https://coeut.iitm.ac.in/ITS_synthesis.pdf

Vatansever, F., & Hamblin, M. R. (2012). Far infrared radiation (FIR): its biological effects and medical applications. *Photonics and Lasers in Medicine, 1*(4), 255-266.

Viola, P., Jones, M., & Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision, 63*(2), 153-161.

Weimer, D., Scholz-Reiter, B., & Shpitalni, M. (2016). Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. *CIRP Annals-Manufacturing Technology*.

Wurfl, T. (2016). *How can I decide the kernel size, output maps and layers of CNN?* Recuperado el 10 de Mayo de 2017, de https://www.quora.com/How-can-I-decide-the-kernel-size-output-maps-and-layers-of-CNN

Xu, F., Liu, X., & Fujimura, K. (2005). Pedestrian detection and tracking with night vision. *IEEE Transactions on Intelligent Transportation Systems, 6*(1), 63-71.

Zeiler, M., & Fergus, R. (Septiembre de 2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 818-833.

Zhang, L., Wu, B., & Nevatia, R. (2007). Pedestrian detection in infrared images based on local shape features. *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 1-8.

Zou, H., Sun, H., & Ji, K. (2012). Real-time infrared pedestrian detection via sparse representation. *Computer Vision in Remote Sensing (CVRS), 2012 International Conference on*, 195-198.

Anexos

Anexo A. Definiciones

A.1 Radiación infrarroja

Lucas en (Lucas, 2015) define a la radiación infrarroja como "un tipo de radiación electromagnética, al igual que las ondas microondas, de radio o ultravioleta. La luz infrarroja es parte del espectro electromagnético con la cual el ser humano experimenta en su vida diaria al usar aparatos como el microondas o el celular, aunque la radiación infrarroja es invisible al ojo humano, este se puede sentir como calor".

La radiación infrarroja al ser considerada una onda electromagnética puede ser expresada en términos de frecuencia o longitud de onda y se extiende desde la luz roja a 700 nm hasta 1 mm (Electromagnetic Radiation, 2001).

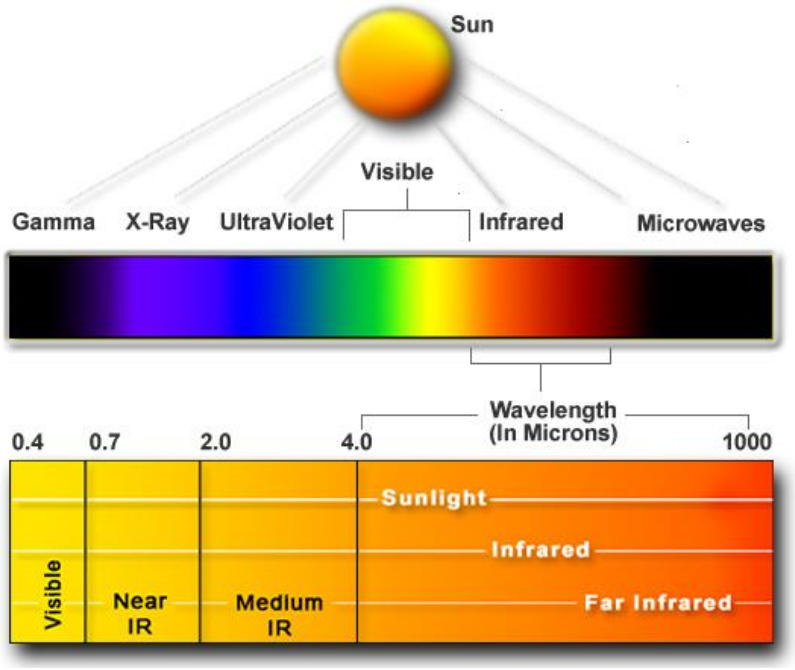


Figura 1. El infrarrojo en el espectro electromagnético.
Fuente: (Infrared Explained, s.f.)

En el año de1800 el astrónomo Sir William Herschel sin proponérselo descubrió la luz infrarroja. En aquel tiempo se sabía que la luz del sol aparte de aportar con la luz visible también aporta con calor, así que Herschel realizo un experimento para

descubrir de donde viene el calor, el cual consistía en medir la temperatura de cada color, para esto descompuso la luz mediante un prisma de vidrio y puso un termómetro en cada color tal como se muestra en la **Figura 2**. Herschel se dio cuenta que la temperatura aumentaba mientras más se acercaba al rojo así que coloco un termómetro justo al lado del rojo donde no había ningún color y se dio cuenta que la temperatura era aún mayor. Entonces Herschel concluyo que existe un tipo de radiación invisible al ojo humano al cual se lo conoció después como radiación infrarroja porque está justo por debajo del color rojo (Experimento de Herschel en la banda infrarroja, s.f.).

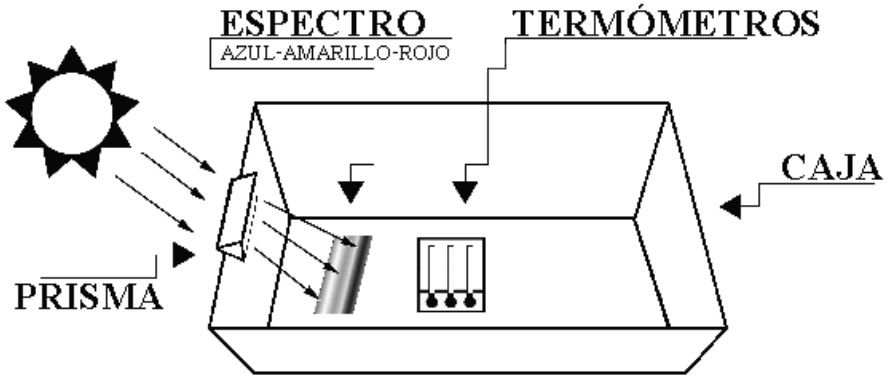


Figura 2. Experimento de Herschel en la banda infrarroja.
Fuente: (Experimento de Herschel en la banda infrarroja, s.f.)

Son varias las instituciones que subdividen al espectro infrarrojo, por ejemplo la Comisión sobre Iluminación (CIE, por sus siglas en francés Commission Internationale de l'Eclairage) presenta tres subdivisiones para la radiación infrarroja:

Tabla 1.
Clasificación de la radiación infrarroja según CIE.

Nombre	**Longitud de onda**	**Energía de fotón (THz)**
Near infrared/IR-A	0.7– 1.4 μm (700– 1400 nm)	215– 430
Mid infrared/IR-B	1.4– 3.0 μm (1400– 3000 nm)	100– 215
Far infrared/IR-C	3.0– 100 μm (3000 nm– 0.1 mm)	3– 100

Fuente: (Vatansever & Hamblin, 2012)**.**

ISO provee otra clasificación alternativa dada por la siguiente tabla:

Tabla 2.
Clasificación de la radiación infrarroja según ISO 20473:2007.

Nombre	**Longitud de onda**
Near-Infrared	780 nm–3 μm
Mid-Infrared	3–50 μm
Far-Infrared	50 μm–1 mm

Fuente: (Robles-Kelly & Huynh, 2013)**.**

Aunque la clasificación más comúnmente usada está dada por el Instituto Alemán de Estandarización (DIN, por sus siglas en alemán Deutsches Institut für Normung) y se define de la siguiente manera:

Tabla 3.
Clasificación de la radiación infrarroja según DIN.

Nombre	**Longitud de onda**
Near infrared, NIR or IR-A	0.75–1.4 micrometers (μm)
Short wavelength (shortwave) IR, SWIR ó IR-B	1.4–3 μm
Mid wavelength IR, MWIR ó IR-C	3–8 μm
Long wavelength IR, LWIR ó IR-C	8–15 μm
Far infrared FIR	15–1,000 μm

Fuente: (Infrared, 2014)**.**

A.2 Sistemas de visión nocturno

Los sistemas de visión nocturno tienen la capacidad de que en condición de baja iluminación permiten al usuario tener una mayor visibilidad de la otorgada por las luces delanteras.

Los sistemas de visión nocturno están basados en tecnología infrarroja cercana o lejana, la cual detecta radiación infrarroja principalmente emitida por peatones o

ciclista y la presentan en una pantalla como si se tratara de un video. Estos sistemas se encuentran instalados en la parte delantera del vehículo, en los luces o en la parrilla (Cisneros, 2008).

Los sistemas de visión nocturna en el infrarrojo cercano cuentan principalmente con los siguientes componentes (Cisneros, 2008):

- Emisores de luz infrarroja.
- Cámara de sensibilidad al infrarrojo.
- Pantalla.

Estos sistemas en vez de usar luz visible para iluminar la carretera utilizan luz infrarroja cercana con una longitud de onda cercana a 0.9 mm para luego con la cámara empezar a capturar las imágenes y transmitirlas en tiempo real a la pantalla (Cisneros, 2008).

Los sistemas de visión nocturna en el infrarrojo lejano cuentan principalmente con:

- Cámara sensible a luz infrarroja lejana.
- Pantalla.

A diferencia del anterior se trata de un tipo de sensado pasivo ya que no requiere ningún tipo de excitación externa, las imágenes se capturan directamente con la cámara infrarroja.

Figura 3. Sistema de visión en el infrarrojo.

Fuente: (Cisneros, 2008)

Uno de los problemas que se tiene al utilizar estos sistemas es que obliga al conductor a estar pendiente de la carretera como de la pantalla, lo que a la larga distrae al conductor o puede causarle fatiga ocular con lo cual la nueva tendencia que se tiene busca proyectar las imágenes que se obtienen de la cámara infrarroja en el parabrisa para fijar la atención del conductor en una sola cosa (¿Qué son los sistemas de visión nocturna?, 2013).

A.3 Tipo de capas en CNN

A.3.1 Capas convolucionales

Esta capa se encarga de realizar la operación de convolución (dada por la ecuación) entre el resultado de la capa anterior con el banco de filtros o kernels que se tenga generando un mapa de activaciones o características. Después de esta capa generalmente se aplica una función de activación al igual que las redes neuronales.

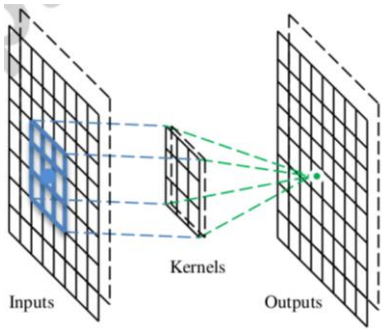


Figura 4. Calculo de la convolución.

Fuente: (Guo, y otros, 2016)

$$y_j = f\left(\sum_i k_{ij} * x_i + b_j\right)$$

Donde:

x_i Representa el canal i de la entrada.

y_j Representa el canal j en el mapa de activaciones.

k_{ij} Representa el kernel.

b Representa el bias.

f Representa la función de activación.

A.3.2 Capas de agrupamiento

Comúnmente se pone una capa de agrupamiento entre dos capas convolucionales para reducir el tamaño espacial de la entrada aplicando alguna operación como el máximo o el promedio en una región tal como se muestra en la **Figura 6**, esta operación se aplica en todos los canales independientemente de la profundidad de la entrada tal como se observa en la **Figura 5** (Karpathy, s.f.).

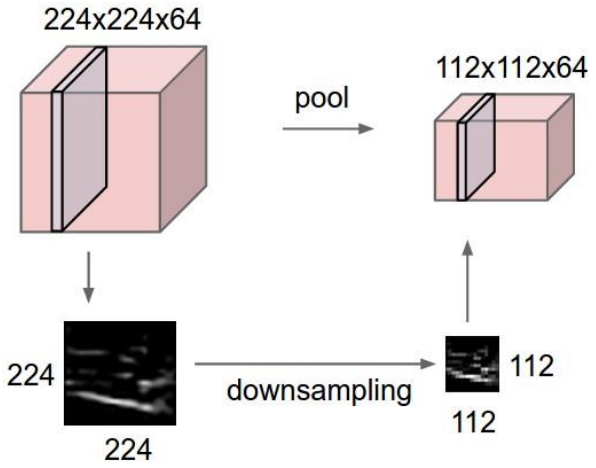


Figura 5. Aplicación de la capa de agrupamiento.

Fuente: (Karpathy, s.f.)

Estudios han comparado el rendimiento de las dos operaciones de agrupamiento más conocidas, el máximo y el promedio y han encontrado que al aplicar el máximo se tiene una convergencia más rápida es por esto que es la más usada, aunque la nueva tendencia planea eliminar esta capa y más bien se plantea reducir la dimensionalidad en la capas convolucionales (Karpathy, s.f.).

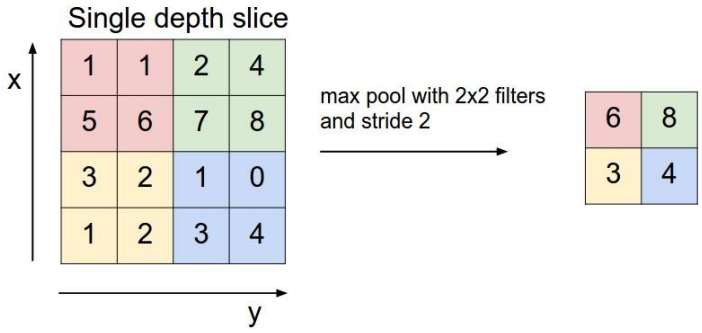


Figura 6. Max-pooling.

Fuente: (Karpathy, s.f.)

$$y_{ijk} = max\ _{p,q}(x_{i,j+p,k+q}) \tag{24}$$

Donde:

y_{ijk} Es el resultado de aplicar la capa de agrupamiento.

$x_{i,j,k}$ Es el valor del canal *i* en la posición *j,k.*

p Es el índice vertical en la vecindad local.

q Es el índice horizontal en la vecindad local.

A.3.3 Capas totalmente conectadas

Esta capa se denomina totalmente conectada debido a que todas las unidades están conectadas entre sí -al igual que las redes neuronales artificiales- a diferencia de las capas convolucionales las cuales comparten parámetros entre sus unidades.

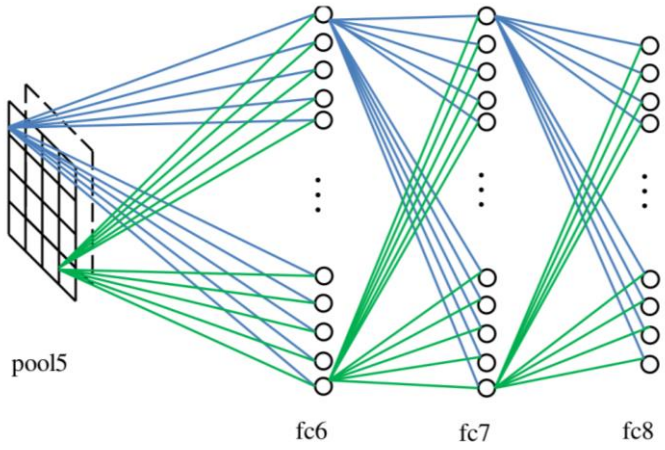


Figura 7. Capa totalmente conectada.
Fuente: (Guo, y otros, 2016)

Las capas totalmente conectadas se entrenan como una red neuronal artificial y abarcan el 90% de los parámetros de un CNN (Guo, y otros, 2016).

A.3.2 Arquitecturas CNN

Existen arquitecturas que han contribuido y han aporta de cierta forma con el desarrollo de CNN las cuales se presentan en la siguiente tabla:

Tabla 4.
Modelos CNN más importantes.

Método	Configuración	Contribución
AlexNet	Cinco capas convolucionales + tres capas totalmente conectadas.	Sirvió de base para las demás.
Clarifai	Cinco capas convolucionales + tres capas totalmente conectadas.	Evalúa la función de las capas de activación intermedias.
SPP	Cinco capas convolucionales + tres capas totalmente conectadas.	Propuesta de la capa SPP la cual elimina la necesidad de mantener constante la resolución de la imagen.
VGG	Trece/Quince capas convolucionales + tres capas totalmente conectadas.	Una evaluación de las redes de mayor profundidad.
GoogleLeNet	Veintiún capas convolucionales + una capa totalmente conectada.	Aumenta la profundidad y el ancho de la red sin aumentar los requisitos computacionales.

Fuente (Guo, y otros, 2016)**.**

A.4 Caffe

Caffe es un sistema de código abierto el cual provee acceso a los algoritmos más importantes del estado del arte en *deep learning*. El código fuente de caffe está escrito en C++ de una forma eficiente y clara. Además Caffe provee enlaces en MATLAB y Python para un desarrollo más rápido. Caffe soporta cálculos acelerados en GPU a través de CUDA. En (Jia, y otros, 2014) dice que usando una GPU K40 o Titan se pueden procesar 40 millones de imágenes por día añadiendo además la ventaja que caffe separa la ejecución de la implementación con lo cual se define en una archivo aparte la arquitectura y en otro las características de entrenamiento.

Una de las ventajas de caffe sobre otros *frameworks* es que su código fuente está escrito en C++ lo cual ayuda a tener una rápida ejecución, la otra ventaja es que en el

caffe model zoo se tiene disponible las principales arquitecturas del estado de arte de redes neuronales convolucionales con lo cual facilita la sintonización de una red con estos modelos pre entrenados.

Tabla 5.
Comparación de diferentes *frameworks* en *deep learning*.

Framework	**Código fuente**	**Enlaces**	**CPU**	**GPU**	**Código abierto**	**Modelos pre entrenados**
Caffe	C++	Python, MATLAB	✓	✓	✓	✓
cuda-convnet	C++	Python		✓	✓	
Decaf	Python		✓		✓	✓
OverFeat	Lua	C++, Python	✓			✓
Theano/Pylearn2	Python		✓	✓	✓	
Torch7	Lua		✓	✓	✓	

Fuente: (Jia, y otros, 2014)**.**

Anexo B. Implementación

B.1 Lectura de datos

Caffe tiene la ventaja que permite definir al usuario sus propias capas, es por esto que para la lectura de datos se creó una nueva capa llamada capa de datos la cual permite leer un archivo de entrenamiento en donde constan todas la imágenes a utilizar en conjunto con los candidatos de la forma como se muestra en la **Figura 8**, en donde la primera línea representa el número de imagen, la segunda línea la ruta de la imagen, las demás líneas representan los candidatos encontrados con el método de generación de candidatos que incluye la clase (peatón o no peatón) esta clase se determina según el grado de solapamiento que se tenga con algún candidato, incluye también el solapamiento y por ultimo las coordenadas del cuadro delimitador.

```
# 0
/ruta img
clase iou x1 y1 x2 y2
clase iou x1 y1 x2 y2
clase iou x1 y1 x2 y2
…….
```

Figura 8. Formato del archivo utilizado en el entrenamiento.

Estas imágenes se leen aleatoriamente en la capa de datos y son alimentadas a la red en cada iteración, una vez que se tiene definida esta capa, se definen las demás capas según la arquitectura determinada en la sección anterior. El archivo utilizado en la definición de la arquitectura está disponible en la sección de anexos.

B.2 Definición Solver

En este archivo se definieron los siguientes parámetros:

Tabla 6.
Parámetros de entrenamiento.

Parámetro	**Valor**
Tasa de aprendizaje	0.002
Factor de castigo	0. 0005
Momentum	0.9
Gamma	0.0001
Power	0.75

Caffe ofrece varios métodos de minimización de la función de costo, el método elegido para resolver la arquitectura en este trabajo es Descenso de Gradiente Estocástico (SGD, por sus siglas en inglés Stochastic Gradient Descent) el cual tiene la ventaja que no calcula la función de costo sobre toda la base de datos de entrenamiento si no esta se calcula solo sobre un minibatch muestreado de la base de datos. La ecuación de actualización de los pesos es la siguiente:

$$V_{t+1} = \mu V_t - \alpha \nabla L(W_t)$$

$$W_{t+1} = W_t + V_{t+1}$$

Donde:

μ Representa el momentum.

α Taza de aprendizaje.

$\nabla L(W_t)$ Función de costo.

El valor de la taza de aprendizaje se actualiza en cada iteración según la ecuación:

$$\alpha = lr * (1 + gamma * iter)^{-power}$$

B.3 Diagrama de clases del sistema implementado

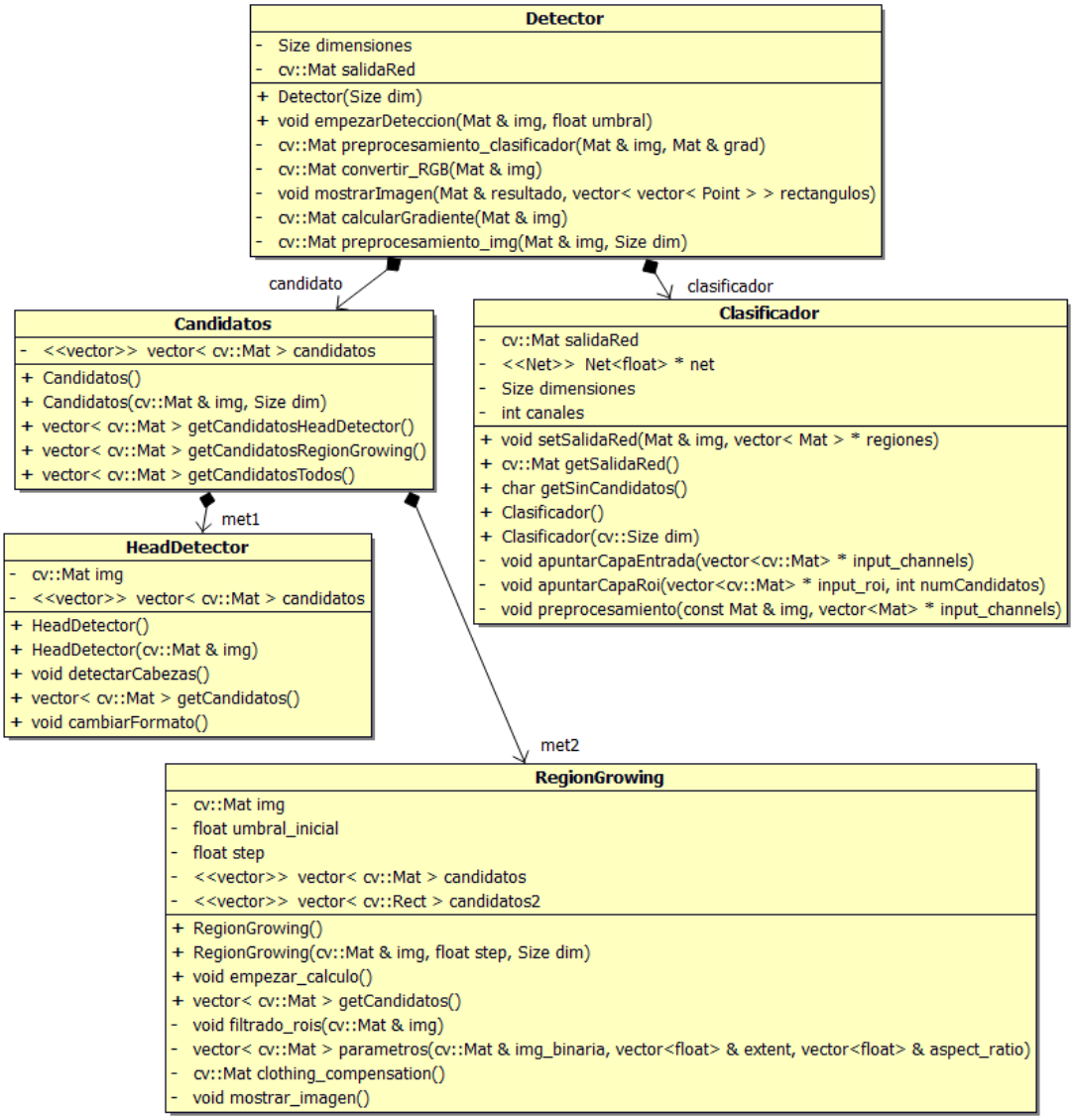


Figura 9. Diagrama de clases del sistema implementado.