



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad de Ingeniería de Sistemas e Informática

Escuela Profesional de Ingeniería de Software

**Detección de situaciones de violencia física
interpersonal en videos usando técnicas de aprendizaje
profundo**

TESIS

Para optar el Título Profesional de Ingeniero de Software

AUTOR

Kent Jhunion CUADROS RAMOS

ASESOR

Dr. Hugo David CALDERÓN VILCA

Lima, Perú

2023



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Cuadros, K. (2023). *Detección de situaciones de violencia física interpersonal en videos usando técnicas de aprendizaje profundo*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniería de Software]. Repositorio institucional Cybertesis UNMSM.

**Metadatos complementarios autor/
asesor**

Datos de autor	
Nombres y apellidos	Kent Jhunior Cuadros Ramos
Tipo de documento de identidad	DNI
Número de documento de identidad	71535381
URL de ORCID	https://orcid.org/0009-0007-9165-3785
Datos de asesor	
Nombres y apellidos	Hugo David Calderón Vilca
Tipo de documento de identidad	DNI
Número de documento de identidad	01317748
URL de ORCID	https://orcid.org/0000-0002-1177-4947
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	José Alfredo Herrera Quispe
Tipo de documento	DNI
Número de documento de identidad	40362859
Miembro del jurado 1	
Nombres y apellidos	Juan Gamarra Moreno
Tipo de documento	DNI
Número de documento de identidad	20039857
Miembro del jurado 2	
Nombres y apellidos	Ciro Rodriguez Rodriguez
Tipo de documento	DNI
Número de documento de identidad	06020241
Datos de investigación	
Línea de investigación	C.0.3.22. Ingeniería de software

Grupo de investigación	Inteligencia Artificial (INTGARTI)
Agencia de financiamiento	Perú. Universidad Nacional Mayor de San Marcos. Vicerrectorado de Investigación y Posgrado. Proyectos de Investigación con Financiamiento para Grupos de Investigación. PCONFIGI. Código: C21201361. Resolución: 005753-2021-R/UNMSM
Ubicación geográfica de la investigación	Universidad Nacional Mayor de San Marcos Edificio: Facultad de Ingeniería de Sistemas e Informática País: Perú Departamento: Lima Provincia: Lima Distrito: Cercado de Lima Calle: Av. Carlos Germán Amezaga 375 Latitud: -12.053697 Longitud: -77.085774
Año o rango de años en que se realizó la investigación	Septiembre 2019 - agosto 2023
URL de disciplinas OCDE	Ciencias de la computación https://purl.org/pe-repo/ocde/ford#1.02.01 Sistemas de automatización, Sistemas de control https://purl.org/pe-repo/ocde/ford#2.02.03 Ingeniería de sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Profesional de Ingeniería de Software

Acta de Sustentación de Tesis

Siendo las diez (10) horas del día diecinueve (19) del mes octubre de año 2023, se reunieron en la sala virtual <https://meet.google.com/cph-apjd-caf> presidido por el Dr. José Alfredo Herrera Quispe (Presidente) y el Mg. Juan Gamarra Moreno (Miembro), para la sustentación de la tesis Intitulada: **"DETECCIÓN DE SITUACIONES DE VIOLENCIA FÍSICA INTERPERSONAL EN VIDEOS USANDO TÉCNICAS DE APRENDIZAJE PROFUNDO"**, por el Bachiller **Kent Jhunion CUADROS RAMOS**, para obtener el Título Profesional de Ingeniero de Software.

Acto seguido de la exposición de la Tesis, el Presidente invitó al Bachiller a responder las preguntas formuladas por los Miembros del Jurado.

El Bachiller, en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez las preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el bachiller obtuvo la nota de 17 (Diecisiete).

A continuación, el Presidente del Jurado Dr. José Alfredo Herrera Quispe, declara al Bachiller Ingeniero de Software.

Siendo las 12:00 horas, se levantó la sesión.

Dr. José Alfredo Herrera Quispe
Presidente

Mg. Juan Gamarra Moreno
Miembro

Dr. Hugo David Calderón Vilca
Asesor

Dr. Ciro Rodríguez Rodríguez
Miembro



CERTIFICADO DE SIMILITUD

Yo **HUGO DAVID CALDERÓN VILCA** en mi condición de asesor acreditado con la **Resolución** **Directoral** **N°000008-2022-EPISW-FISI/UNMSM** de la **tesis/monografía/informe** de investigación/trabajo académico, cuyo título es **DETECCIÓN DE SITUACIONES DE VIOLENCIA FÍSICA INTERPERSONAL EN VIDEOS USANDO TÉCNICAS DE APRENDIZAJE PROFUNDO**, presentado por el bachiller Kent Jhunion **CUADROS RAMOS**, para optar el grado/título de **Ingeniero de Software**. CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante el software de similitud textual, el documento evaluado cuenta con el porcentaje de **19%** de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

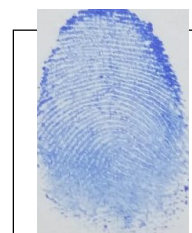
Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del grado/ **título**/ especialidad correspondiente.

Firma del Asesor _____

DNI: 01317748

Dr. Hugo David Calderón Vilca

Asesor



AGRADECIMIENTOS

A Dr. Calderón Vilca, Hugo David por su orientación, dedicación, consejos y revisiones para que este trabajo cumpla con los objetivos trazados. A todas aquellas personas que indirectamente ayudaron para culminar este trabajo y que muchas veces constituyen un invalorable apoyo. Y por encima de todo damos gracias a Dios.

RESUMEN

Debido al aumento de inseguridad en el país y como medida preventiva se buscó reforzar el sistema de videovigilancia, donde se enfocó en la necesidad de integrar nuevas tecnologías para supervisar la seguridad ciudadana como es el caso del uso de la visión artificial. Se diseñó una arquitectura con el modelo de red neuronal convolucional Xception y LSTM para la detección de violencia física interpersonal en los videos de sistemas de vigilancia, para el entrenamiento, validación y prueba de la arquitectura del modelo propuesto se utilizó los conjuntos de datos Hockey Fight Dataset y Real Life Violence Situations Dataset, donde los resultados obtenidos en la exactitud de nuestra propuesta en el conjunto de datos Hockey Fight Dataset superó a todos los demás métodos, en el caso del conjunto de datos Real Life Violence Situations Dataset que cuenta 2000 videos en contraste de otros conjuntos de datos utilizados para la detección de violencia, se obtuvieron buenos resultados en la exactitud mayores al 90%.

Palabras Claves: Violencia física interpersonal, Detección, Redes Neuronales Convolucionales, Redes Neuronales Recurrentes.

ABSTRACT

Due to the increase in insecurity in the country and as a preventive measure, it seeks to reinforce the video surveillance system, where it will focus on the need to integrate new technologies to supervise citizen security, such as the use of artificial vision. An architecture will be designed with the Xception and LSTM convolutional neural network model for the detection of interpersonal physical violence in the videos of surveillance systems, for training, validation and testing of the architecture of the proposed model we use the Hockey Fight Dataset data sets and Real Life Violence Situations Dataset, where the results obtained in the accuracy of our proposal in the Hockey Fight Dataset will outperform all other methods, in the case of the Real Life Violence Situations Dataset that counts 2000 videos In contrast to other data sets used for violence detection, it will yield good results in accuracy greater than 90%.

Keywords: Interpersonal physical violence, Detection, Convolutional Neural Networks, Recurrent Neural Networks.

ÍNDICE

1.	Capítulo I: Introducción	7
1.1.	Antecedentes del problema	7
1.1.1.	Violencia en el mundo	7
1.1.2.	Violencia y seguridad ciudadana en Latinoamérica y el Caribe	8
1.1.3.	Percepción de inseguridad ciudadana en Perú	9
1.2.	Descripción del problema	12
1.2.1.	Formulación del problema	13
1.3.	Importancia	14
1.4.	Motivación	16
1.5.	Objetivos	18
1.5.1.	Objetivo general	18
1.5.2.	Objetivos específicos	18
1.6.	Propuesta	19
1.7.	Estructura de Tesis	19
2.	Capítulo II: Marco teórico	20
2.1.	Redes neuronales artificiales	20
2.1.1.	Aprendizaje supervisado	21
2.1.2.	Aprendizaje no supervisado	21
2.2.	Redes neuronales convolucionales	22
2.3.	Redes neuronales recurrentes	23
2.3.1.	Redes de memoria a corto y largo plazo	24
3.	Capítulo III: Estado del arte	25
3.1.	Modelos para la detección de eventos	28
3.2.	Métodos de detección en videos	33
3.3.	Aplicaciones en Sistemas de Vigilancia	45
4.	Capítulo IV: Metodología de la investigación	55
4.1.	Dataset	55
4.2.	Componentes	56
4.2.1.	Red neuronal convolucional Xception	56
4.2.2.	Long short-term memory (LSTM)	57
4.2.3.	Capas totalmente conectadas	58
4.3.	Flujo de trabajo de entrenamiento del modelo de red neuronal propuesto para la detectar violencia física interpersonal en videos.	59
4.3.1.	Preparación del Dataset de videos	60

4.3.2.	Preprocesamiento de los videos	61
4.3.3.	Entrenamiento del modelo de red neuronal	65
4.4.	Metodología de desarrollo e implementación del prototipo que detecta violencia física interpersonal.....	65
4.4.1.	Descripción general del prototipo.....	65
4.4.2.	Arquitectura de contexto del prototipo	66
4.4.3.	Arquitectura Lógica	67
4.4.4.	Arquitectura de componentes	69
4.4.5.	Arquitectura de bajo nivel.....	71
4.4.6.	Arquitectura física.....	71
5.	Capítulo V: Resultados y discusión	74
5.1.	Resultados	74
5.1.1.	Dataset.....	74
5.1.2.	Traza del entrenamiento y métricas	76
5.1.3.	Prueba del modelo y métricas	79
5.2.	Discusión.....	81
5.3.	Conclusiones	84
5.4.	Trabajos futuros	85
	Referencias.....	86

ÍNDICE DE TABLAS

Tabla 1 Cifras estimadas de defunciones debidas a la violencia en el mundo (Garmendia Lorena, 2011)	7
Tabla 2 Clasificación de la violencia de acuerdo con la OMS (Garmendia Lorena, 2011).....	8
Tabla 3 Resumen del Estado del Arte	25
Tabla 4 Traza de entrenamiento	76
Tabla 5 Comparación de resultados de clasificación	81

ÍNDICE DE FIGURAS

Figura 1 Tasa de homicidios cada 100.000 habitantes (Beliz, 2012)	9
Figura 2 Percepción de inseguridad según ámbitos de estudio (Quispe Huamaní et al., 2018)..	10
Figura 3 Percepción de inseguridad según ámbitos geográficos de Lima Metropolitana (Quispe Huamaní et al., 2018).....	11
Figura 4 Percepción de inseguridad según áreas interdistritales de la Provincia de Lima (Quispe Huamaní et al., 2018).....	12
Figura 5 Percepción de inseguridad por existencia de vigilancia su zona o barrio (Condori Huahuachampi et al., 2018)	14
Figura 6 Problema más importante de un país en América Latina (Latinobarómetro, 2017)	15
Figura 7 Problemas principales del Perú (PUCP, 2016).....	16
Figura 8 Población que manifiesta existencia de vigilancia (Durand Carrión et al., 2017)	16
Figura 9 Esquema de una Neurona Artificial	21
Figura 10 Arquitectura General de una Red Neuronal Convolutacional.....	23
Figura 11 Arquitectura de la unidad RNN.....	24
Figura 12 Arquitectura de la unidad LSTM.....	24
Figura 13 Real Life Violence Dataset.....	56
Figura 14 Arquitectura de CNN Xception.....	57
Figura 15 Arquitectura LSTM	58
Figura 16 Arquitectura de Capas Totalmente Conectadas.....	59
Figura 17 Diagrama de flujo de trabajo de entrenamiento del modelo de red neuronal.....	60
Figura 18 Secuencias de imágenes preprocesadas y extraídas del video	62
Figura 19 Características extraídas de una imagen en la primera capa de CNN Xception	63

Figura 20 Características extraídas de una imagen en la capa intermedia avg_pool de CNN

Xception..... 65

Figura 21 Diagrama de contexto del prototipo 66

Figura 22 Diagrama de contenedores del prototipo Violence Detection..... 68

Figura 23 Diagrama de componentes 70

Figura 24 Diagrama de clases del dominio violencia 71

Figura 25 Diagrama de despliegue 72

Figura 26 Muestra de videos sin presencia de violencia del conjunto de datos 75

Figura 27 Muestra de videos con presencia de violencia del conjunto de datos 76

Figura 28 Gráfico de precisión del Modelo 78

Figura 29 Gráfico de pérdidas del Modelo 79

Figura 30 Matriz de confusión..... 80

Capítulo I: Introducción

1.1. Antecedentes del problema

La violencia es una situación donde alguien con mayor poder abusa de otra persona con menor poder, siendo este estudio enfocado en la violencia del tipo físico, donde se observó su importancia dentro del contexto mundial como localmente en el país y la capital.

(Carrión M., 2005) ha afirmado lo siguiente:

Si bien la violencia urbana existe desde que existe la ciudad, es difícil desconocer que, durante estos últimos años, se ha convertido en uno de los temas más importantes de la urbe andina contemporánea, debido, entre otras cosas, a las nuevas formas que ha asumido, a los impactos sociales y económicos y al incremento alarmante de su magnitud. Las violencias se han extendido en todos los países y ciudades de la región, pero con peculiaridades y ritmos de intensidad propios a cada urbe y cada cultura. (pág. 30)

1.1.1. Violencia en el mundo

En Latinoamérica, la violencia es un problema importante, siendo una de las principales causas de muerte, la OMS nos muestra en la Tabla 1, en el informe del año 2000, que las muertes alcanzadas llegaron a la cantidad de 1 659 000, reportándose el 90% en los países de ingresos bajos a medianos. (Garmendia Lorena, 2011)

Tabla 1 Cifras estimadas de defunciones debidas a la violencia en el mundo (Garmendia Lorena, 2011)

Tipo de violencia	Número	Tasa por 100 000 habitantes	Proporción total (%)
Homicidios	520 000	8,8	31,3
Suicidios	815 000	14,5	49,1
Causadas por acciones bélicas	310 000	5,2	18,6

Total	1 659 000	28,8	100,0
Países de ingreso bajo a mediano	1 510 000	32,1	91,1
Países de ingreso alto	149 000	14,4	8,9

En la tabla 2, se encuentra la clasificación de la violencia de acuerdo con la OMS.

Tabla 2 Clasificación de la violencia de acuerdo con la OMS (Garmendia Lorena, 2011)

1. Autoinfligida
1. Comportamiento suicida
2. Autolesiones
2. Interpersonal
1. Familia/Pareja
1. Menores
2. Pareja
3. Anciano
2. Comunidad
1. Amistades
2. Extraños
3. Colectiva
1. Social
2. Política
3. Económica

1.1.2. Violencia y seguridad ciudadana en Latinoamérica y el Caribe

En Latinoamérica y el Caribe, la situación de inseguridad ciudadana tiene un marcado deterioro sin importar el punto de vista que se dé, ya sea de manera objetiva como subjetiva, entonces claramente la seguridad es un problema de desarrollo. (Beliz, 2012)

- En el Gráfico 1 vemos que la región de América latina y el caribe es una de las más violentas del mundo. Según las últimas estadísticas disponibles, los Países Andinos exhiben una tasa de 24,8 homicidios por cada 100.000 habitantes, la cual es tres veces superior al promedio mundial (8 homicidios cada 100.000 habitantes). Centroamérica es la subregión más problemática en este sentido, con tasas que llegan a los 33,6 homicidios

cada 100.000 habitantes, 5 mientras que el Cono Sur presenta las tasas más bajas (10,9 homicidios cada 100.000 habitantes).

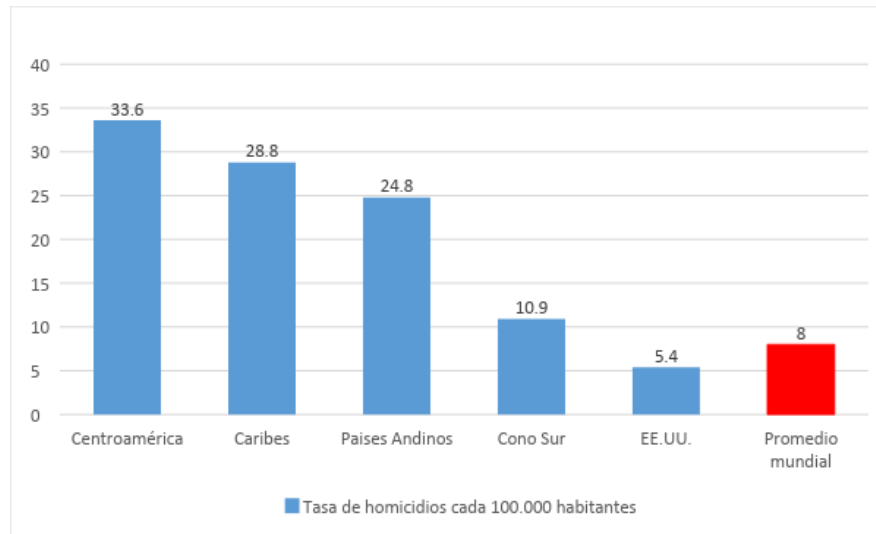


Figura 1 Tasa de homicidios cada 100.000 habitantes (Beliz, 2012)

1.1.3. Percepción de inseguridad ciudadana en Perú

(Quispe Huamaní et al., 2018) describe la percepción de inseguridad ciudadana en el Perú, cuando las personas perciben que serán víctimas de alguna manera donde se atente en contra de su seguridad en los próximos doce meses. Estas personas estarán en el área urbana y tienen de 15 años a más.

- **Nacional Urbano**

En el 2017 la percepción de inseguridad ciudadana a nivel nacional urbano alcanzo el 86.9% de la población, de igual forma, en ciudades de 20 mil a más habitantes el porcentaje fue de 88,7%, en tanto que en los centros poblados entre 2 mil y menos de 20 mil habitantes fue de 82,0%. (Quispe Huamaní et al., 2018)

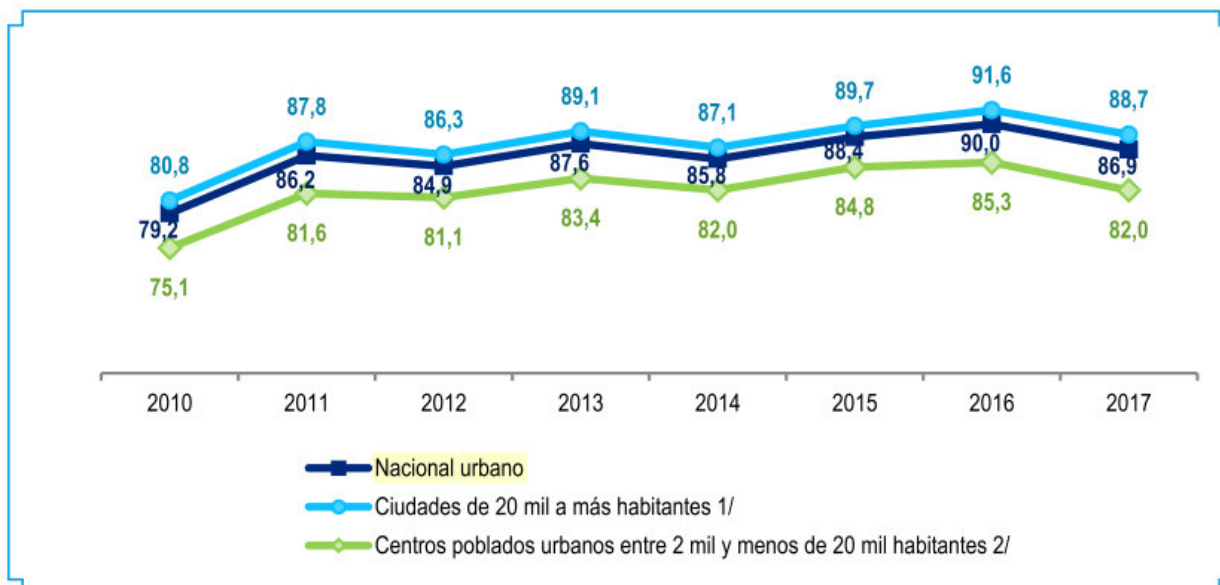


Figura 2 Percepción de inseguridad según ámbitos de estudio (Quispe Huamaní et al., 2018)

- **Lima Metropolitana**

La percepción de inseguridad ciudadana a nivel de Lima Metropolitana alcanzo el 90.0% de la población, de igual forma, a nivel de Provincia de Lima el porcentaje fue de 90,3%, mientras que en la Provincia Constitucional del Callao presento el menor porcentaje con el 87,9%. (Quispe Huamaní et al., 2018)

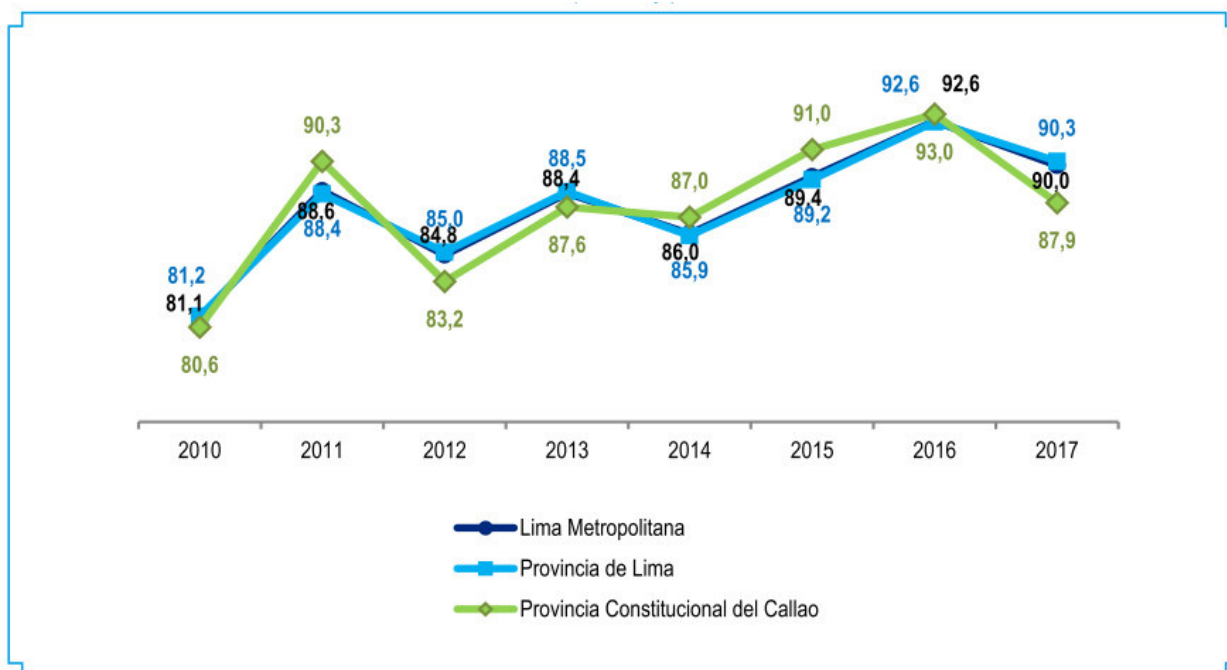


Figura 3 Percepción de inseguridad según ámbitos geográficos de Lima Metropolitana
(Quispe Huamaní et al., 2018)

- **Áreas interdistritales de la provincia de Lima**

Los resultados sobre la percepción de inseguridad ciudadana, según áreas interdistritales de la Provincia de Lima, alcanzaron el 92,5%, 91,0% ,89,3% y 89,0% en el área Lima Sur, Lima Norte, Lima Este y Lima Centro respectivamente. (Quispe Huamaní et al., 2018)

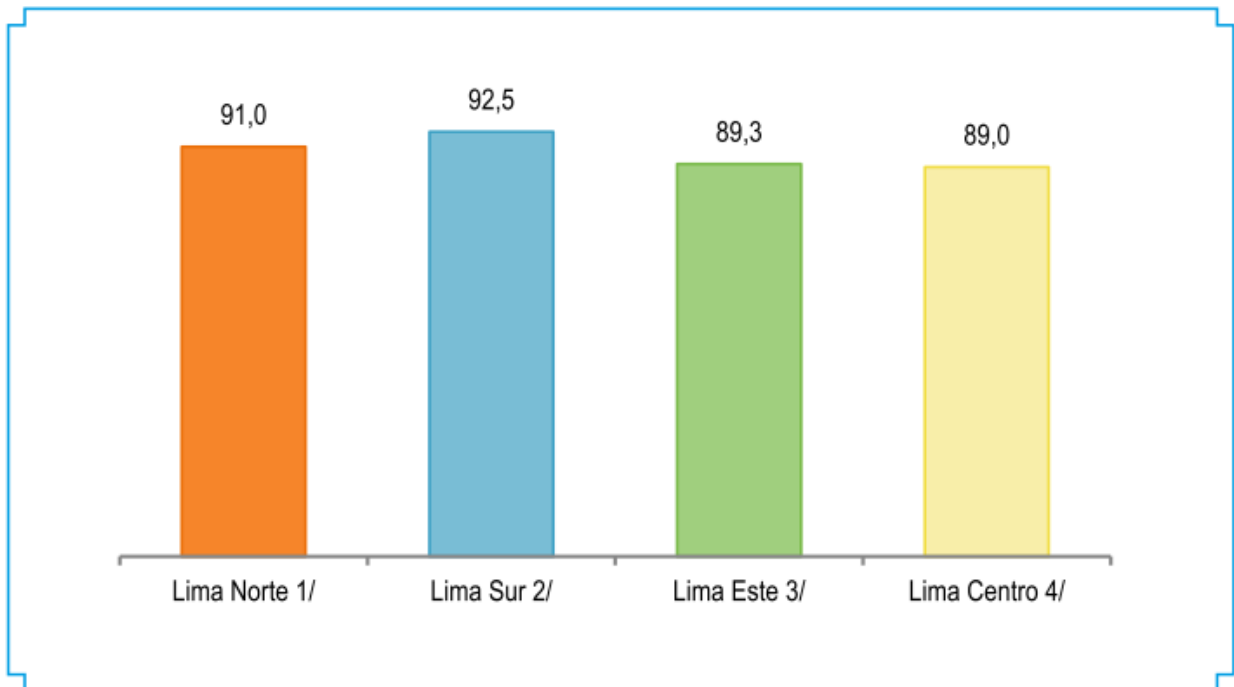


***Figura 4** Percepción de inseguridad según áreas interdistritales de la Provincia de Lima*

(Quispe Huamaní et al., 2018)

1.2. Descripción del problema

En (Reyna & Toche, 1999) nos indica que el alto grado de inseguridad generado por la violencia y la delincuencia, en Perú como en toda América Latina, impiden el crecimiento económico y la disminución de la pobreza.

Podemos definir entonces el problema de la inseguridad ciudadana como el miedo y/o la percepción a posibles agresiones, asaltos, secuestros, violaciones, de los cuales podemos ser víctimas, teniendo una relación de manera directa con la violencia física interpersonal. Entonces el problema de la inseguridad ciudadana relacionada con la violencia física interpersonal viene siendo de interés social como las personas y la misma sociedad donde se interrelacionan, en consecuencia, siendo de interés gubernamental, pues las consecuencias son de orden social.

De (Reyna & Toche, 1999) podemos interpretar que en la actualidad el problema anteriormente mencionado, es una de las principales características de todas las sociedades que cuentan o no con un sistema de vigilancia buscando una mejor vigilancia para la seguridad de sus habitantes, y es que vivimos en un mundo en el que la extensión de la violencia se ha desbordado en un clima generalizado de criminalidad.

1.2.1. Formulación del problema

- **Problema general**

¿Se puede implementar un prototipo de detección de situaciones de violencia física interpersonal en videos usando técnicas de aprendizaje profundo?

- **Problemas específicos**

PE1: ¿Se logrará identificar un conjunto de datos de videos con situaciones de violencia en la vida real, para entrenar el complejo modelo de red neuronal convolucional (CNN) con una red neuronal recurrente (LSTM)?

PE2: ¿Es posible diseñar la arquitectura de un complejo modelo de red neuronal que integre una red neuronal convolucional (CNN) con una red neuronal recurrente (LSTM), para la detección de violencia física interpersonal en videos?

PE3: ¿Se puede desarrollar un prototipo de detección de situaciones de violencia física interpersonal en videos usando técnicas de aprendizaje profundo?

PE4: ¿Cómo podemos validar la arquitectura del complejo modelo de red neuronal convolucional (CNN) con una red neuronal recurrente (LSTM) para la detección de violencia física interpersonal en videos?

1.3. Importancia

Debido al aumento de inseguridad en el país y como medida preventiva se busca reforzar el sistema de videovigilancia, donde se enfocará en la necesidad de integrar nuevas tecnologías para supervisar la seguridad ciudadana como es el caso del uso de la visión artificial aplicado para la detección de violencia física interpersonal en videos de seguridad.

- Del grafico de (Condori Huahuachampi et al., 2018) podemos entender que la percepción de inseguridad, de la población que cuentan con vigilancia en su zona o barrio, es menor respecto a las que no cuentan con vigilancia.

Semestre móvil	Percepción de inseguridad	
	Zona o barrio sin vigilancia	Zona o barrio con vigilancia
Indicadores semestrales		
Dic 2016 - May 2017	88,7	87,4
Ene 2017 - Jun 2017	87,8	87,3
Feb 2017 - Jul 2017	87,7	87,2
Mar 2017 - Ago 2017	87,4	86,0
Abr 2017 - Sep 2017	87,4	86,0
May 2017 - Oct 2017	87,4	85,7
Jun 2017 - Nov 2017	87,0	85,3
Jul 2017 - Dic 2017	86,7	84,7
Ago 2017 - Ene 2018	86,5	84,7
Sep 2017 - Feb 2018	86,0	84,9
Oct 2017 - Mar 2018	86,0	84,9
Nov 2017 - Abr 2018	85,8	84,3
Dic 2017 - May 2018	86,2	84,6
Ene 2018 - Jun 2018	85,8	85,0
Diferencia con semestre móvil anterior (puntos porcentuales)		
Dic 2017 - May 2018/ Ene 2018 - Jun 2018	-0,4	0,4
Diferencia con semestre similar del año anterior (puntos porcentuales)		
Ene 2017 - Jun 2017/ Ene 2018 - Jun 2018	-2,0	-2,3

Fuente: Instituto Nacional de Estadística e Informática – Encuesta Nacional de Programas Presupuestales 2016-2018 (información preliminar).

Figura 5 Percepción de inseguridad por existencia de vigilancia su zona o barrio
(Condori Huahuachampi et al., 2018)

- En el gráfico de (Latinobarómetro, 2017) vemos que uno de los problemas más importantes en la región de Latinoamérica, solo superado por el problema de carácter económico, es la inseguridad y la delincuencia.



Figura 6 Problema más importante de un país en América Latina (Latinobarómetro, 2017)

- En (PUCP, 2016) observamos que el 80% de la población peruana considera que el problema principal del país es la delincuencia y la inseguridad ciudadana.



Figura 7 Problemas principales del Perú (PUCP, 2016)

- Y en (Durand Carrión et al., 2017) se interpreta que “El 45,7% de la población de 15 y más años de edad del área urbana manifestó la existencia de vigilancia en su zona o barrio en el año 2017; disminuyó 4,6 puntos porcentuales con relación al año 2011.”

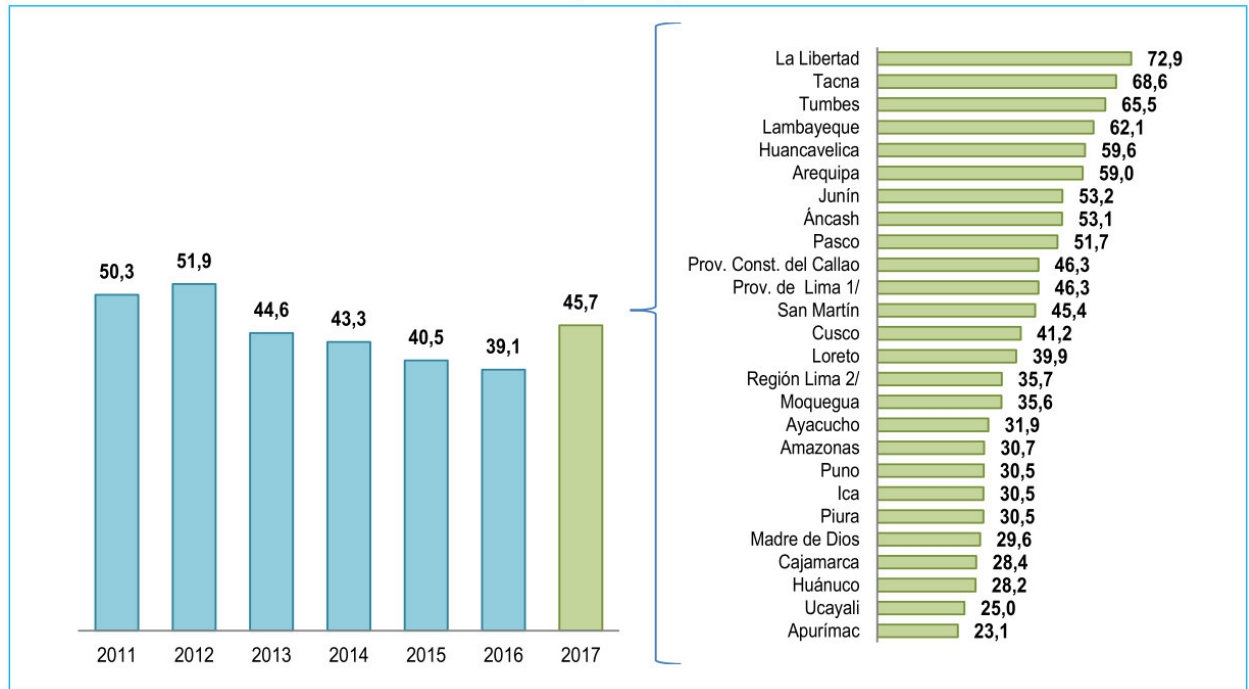


Figura 8 Población que manifiesta existencia de vigilancia (Durand Carrión et al., 2017)

1.4. Motivación

Se han analizado y revisado diversos trabajos de investigación relacionados con la detección de violencia física interpersonal en videos de seguridad ciudadana. En (Roa et al., 2018) proponen un componente de software prototipo para la detección automática de situaciones de inseguridad basadas en el reconocimiento de voz, el procesamiento del lenguaje natural y el reconocimiento de patrones, este componente proporciona un servicio de vigilancia automática que permite detectar en tiempo real las situaciones de inseguridad, generar alertas de seguridad para ayudar a

mitigar dichas situaciones e informar y/o registrar los eventos ocurridos, pero sólo obtienen buenos resultados en ambientes controlados y no en entornos reales debido a la baja calidad de audio y ruidos que se presentan cuando se realizan las grabaciones. En (Izadi et al., 2018) proponen un sistema para la detección de eventos anormales de videos de vigilancia, centrado en un algoritmo de visión de computadora no supervisado en escenas dinámicas, incluyendo escenas de áreas abarrotadas o con poca gente, sin embargo, en este trabajo no consideraron centrarse en extraer características más discriminativas que tiene el ser humano, tampoco la detección de violencia. También se analizó y revisó (Susarla et al., 2018) donde se propone para la detección de humanos armados un sistema de videovigilancia con un reconocedor pre-entrenado, un detector y un rastreador comunes para todos los objetos de vigilancia, para tratar todos ellos simultáneamente siguieron un enfoque de Interacción Humano-Objeto HOI, el sistema ha utilizado el método de redes neuronales recurrentes estructurales (SRNN), para aprender automáticamente tales interacciones, también hacen uso de datos RGB en lugar de RGB-D, pero, sin embargo los métodos actuales de segmentación temporal que utilizan no brindan una gran precisión, esta precisión se podría lograr con segmentación temporal automática utilizando métodos de aprendizaje profundo. En (Bouachir et al., 2018) para la detección de intentos de suicidios proponen un sistema de videovigilancia inteligente que funciona utilizando flujos de profundidad proporcionados por una cámara RGB-D, independientemente de las condiciones de iluminación, este sistema está basado en la visión para detección automática del suicidio por ahorcamiento, apoyado en el Algoritmo 1 (Bouachir et al., 2018) para la estimación de parámetros de escala y el Algoritmo 2 (Bouachir et al., 2018) en el reconocimiento de acciones, también proponen el balanceo del conjunto de datos desequilibrado, pero en la evaluación del sistema se usaron datos simulados, lo cual no nos brinda la certeza que realmente funcionara con datos que se obtienen en un ambiente

real. En (Ben Mabrouk & Zagrouba, 2017) (T. Zhang et al., 2018) (Sudhakaran & Lanz, 2017) para el entrenamiento de sus modelos neuronales profundos usaron los conjuntos de datos Hockey Fight Dataset, The Violent Flows dataset, Movies Dataset, The BEHAVE dataset, The Crowd Violence dataset, pero, sin embargo estos conjuntos de datos de videos analizados no están relacionados de manera tan directa con situaciones de violencia en la vida real, la cual se capturan y se monitorean realmente en un sistema de videovigilancia para la seguridad ciudadana.

1.5. Objetivos

1.5.1. Objetivo general

Implementar un prototipo de detección de situaciones de violencia física interpersonal en videos usando técnicas de aprendizaje profundo.

1.5.2. Objetivos específicos

OE1: Identificar un conjunto de datos de videos con situaciones de violencia en la vida real, para entrenar el complejo modelo de red neuronal convolucional (CNN) con una red neuronal recurrente (LSTM).

OE2: Diseñar la arquitectura de un complejo modelo de red neuronal que integre una red neuronal convolucional (CNN) con una red neuronal recurrente (LSTM), para la detección de violencia física interpersonal en videos.

OE3: Desarrollar un prototipo de detección de situaciones de violencia física interpersonal en videos usando técnicas de aprendizaje profundo.

OE4: Validar la arquitectura del complejo modelo de red neuronal convolucional (CNN) con una red neuronal recurrente (LSTM) para la detección de violencia física interpersonal en videos.

1.6. Propuesta

Proponemos un prototipo que permita detectar situaciones de violencia física interpersonal en videos usando técnicas de aprendizaje profundo.

1.7. Estructura de Tesis

La presente tesis está estructurada en 5 capítulos que se detallan a continuación.

- El capítulo 1 en el cual se detallaron los antecedentes, la problemática, la importancia, motivación y los objetivos para esta investigación.
- En el capítulo 2, se describe el marco teórico en el que se define qué son las redes neuronales artificiales y sus tipos de aprendizajes, las redes neuronales convolucionales y las redes neuronales.
- En el capítulo 3, se hace un detalle y análisis del estado del arte de investigaciones que buscan detectar y/o clasificar movimientos, acciones del ser humano principalmente relacionados con la violencia.
- En el capítulo 4, se presenta el aporte en el cual se detalla el dataset a utilizar, el modelo propuesto para detectar situaciones violencia física interpersonal, además de los detalles de implementación del prototipo que detecta violencia.
- En el capítulo 5 se analiza los resultados obtenidos del entrenamiento del complejo modelo de red neuronal convolucional con LSTM, además de presentar las conclusiones y trabajos futuros.

Capítulo II: Marco teórico

Este capítulo describe los conceptos generales de las redes neuronales, de ellas, las redes convolucionales y las recurrentes, las cuales son usadas en el procesamiento de video y de secuencias de imágenes.

2.1. Redes neuronales artificiales

Una red neuronal artificial (ANN) es un esquema de computación distribuida inspirada en la estructura del sistema nervioso de los seres humanos. La arquitectura de una red neuronal es formada conectando múltiples procesadores elementales, siendo éste un sistema adaptivo que posee un algoritmo para ajustar sus pesos (parámetros libres) para alcanzar los requerimientos de desempeño del problema basado en muestras representativas (Fausett, 1994).

Por lo tanto, podemos señalar que una ANN es un sistema de computación distribuida caracterizada por:

- Un conjunto de unidades elementales, cada una de las cuales posee bajas capacidades de procesamiento.
- Una densa estructura interconectada usando enlaces ponderados.
- Parámetros libres que deben ser ajustados para satisfacer los requerimientos de desempeño.
- Un alto grado de paralelismo.

Es importante señalar que la propiedad más importante de las redes neuronales artificiales es su capacidad de aprender a partir de un conjunto de patrones de entrenamientos, es decir, es capaz de encontrar un modelo que ajuste los datos. El proceso de aprendizaje también conocido como entrenamiento de la red puede ser supervisado o no supervisado (Fausett, 1994).

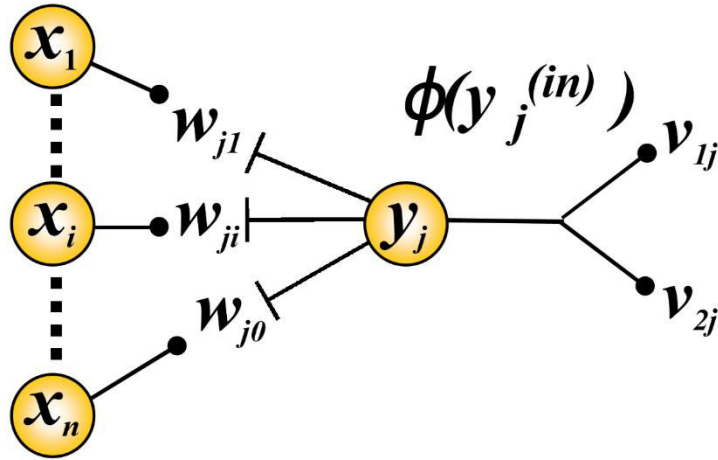


Figura 9 Esquema de una Neurona Artificial

2.1.1. Aprendizaje supervisado

El aprendizaje supervisado consiste en entrenar la red a partir de un conjunto de datos o patrones de entrenamiento compuesto por patrones de entrada y salida. El objetivo del algoritmo de aprendizaje es ajustar los pesos de la red w de manera tal que la salida generada por la ANN sea lo más cercanamente posible a la verdadera salida dada una cierta entrada. Es decir, la red neuronal trata de encontrar un modelo al proceso desconocido que generó la salida y . Este aprendizaje se llama supervisado pues se conoce el patrón de salida el cual hace el papel de supervisor de la red (Haykin, 1998).

2.1.2. Aprendizaje no supervisado

En el aprendizaje no supervisado se presenta sólo un conjunto de patrones a la ANN, y el objetivo del algoritmo de aprendizaje es ajustar los pesos de la red de manera tal que la red encuentre alguna estructura o configuración presente en los datos (Haykin, 1998).

2.2. Redes neuronales convolucionales

Un tipo de redes neuronales, que es el utilizado en este trabajo, son las redes neuronales convolucionales, sus capas en lugar de tener pesos independientes por cada neurona, los pesos se utilizan para definir un filtro convolucional que se aplica a la información de entrada para a continuación, aplicar la función de activación al resultado de la convolución (Ahmed et al., 2019).

En una capa convolucional, hay varios parámetros a definir: El primero es el tamaño del filtro convolucional, seguido del paso de convolución (o stride) que define el desplazamiento que aplicamos al filtro a la hora de realizar la convolución, y finalmente definir como operaremos con los bordes, si rellenaremos la parte exterior de la imagen de entrada para que el filtro pueda operar en esa zona, o si elegiremos que el filtro tan solo opere dentro de la zona de la imagen (se denomina padding a este proceso). Según lo que escojamos en relación a estos parámetros, definiremos la dimensión que tendrá la información a la salida de la capa. Ya en la parte final de la red neuronal convolucional, nos solemos encontrar una capa fully connected, seguida de una capa que transforma los valores a probabilidad (Ahmed et al., 2019).

Las capas fully connected, tiene la característica que todas las neuronas están interconectadas con las neuronas de la capa anterior y posterior. Esta característica no las hace especialmente útiles cuando se trabaja con una gran cantidad de nodos, ya que las conexiones aumentan con mucha rapidez (Ahmed et al., 2019).

Finalmente, hay que aplicar una última capa, que transforme los valores de salida de la capa fully connected en valores probabilísticos, de manera que tengamos un valor adecuado para poder realizar la clasificación. Ejemplos de estas capas podrían ser la capa softmax, que recibiendo la misma cantidad de neuronas que clases, asigna una probabilidad a cada una de las clases en función de los valores recibidos y que definida por la siguiente función mostrada a continuación, o una

capa sigmoid, muy utilizada porque sus valores de salida van de 0 a 1, coincidiendo con los valores posibles de la probabilidad (Sudhakaran & Lanz, 2017).

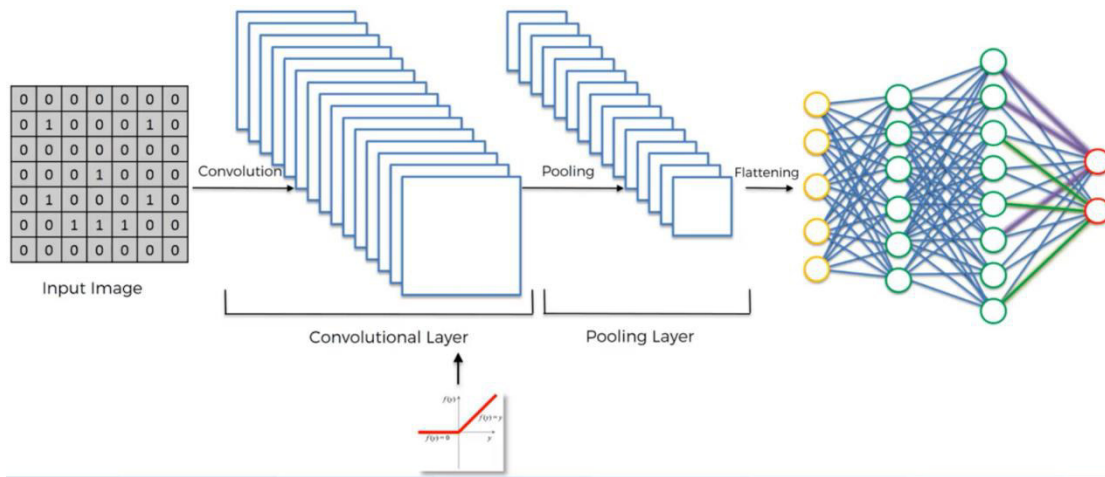


***Figura 10** Arquitectura General de una Red Neuronal Convolucional*

2.3. Redes neuronales recurrentes

Red neuronal recurrente (RNN) es una clase de red neuronal artificial donde las conexiones entre nodos forman un grafo dirigido a lo largo de una secuencia temporal. Esto le permite exhibir un comportamiento dinámico temporal. Los RNN usan su estado interno (memoria) para procesar secuencias de entradas. Esto los hace aplicables a tareas como el reconocimiento de escritura a mano no segmentado o el reconocimiento de voz (Susarla et al., 2018).

La idea detrás de los RNN es hacer uso de información secuencial. En una red neuronal tradicional, todas las entradas (y salidas) son independientes entre sí, pero para muchas tareas es una muy mala idea. Los RNN se denominan recurrentes porque realizan la misma tarea para cada elemento de una secuencia, y la salida depende de los cálculos anteriores. Otra forma de pensar acerca de los RNN es que tienen una "memoria" que captura información sobre lo que se ha calculado hasta ahora (Susarla et al., 2018).

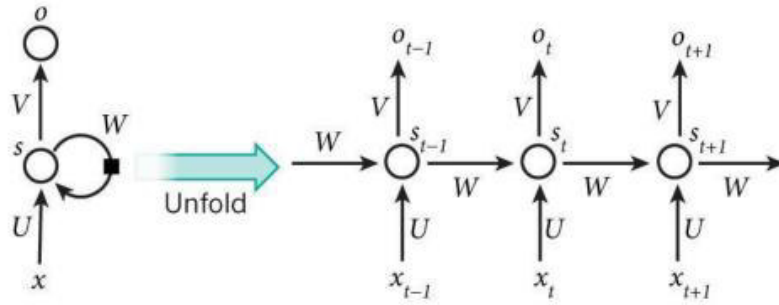


Figura 11 Arquitectura de la unidad RNN

2.3.1. Redes de memoria a corto y largo plazo

Las redes de memoria a corto y largo plazo, generalmente llamadas "LSTM", son un tipo especial de RNN, capaces de aprender dependencias a largo plazo. Fueron introducidos por Hochreiter y Schmidhuber (1997). Los LSTM están diseñados explícitamente para evitar el problema de dependencia a largo plazo (Sudhakaran & Lanz, 2017).

Recordar información durante largos períodos de tiempo es prácticamente su comportamiento predeterminado, no es algo que les cuesta aprender. Todas las redes neuronales recurrentes tienen la forma de una cadena de módulos repetitivos de red neuronal. En los RNN estándar, este módulo repetitivo tendrá una estructura muy simple, como una sola capa de tanh (Sudhakaran & Lanz, 2017).

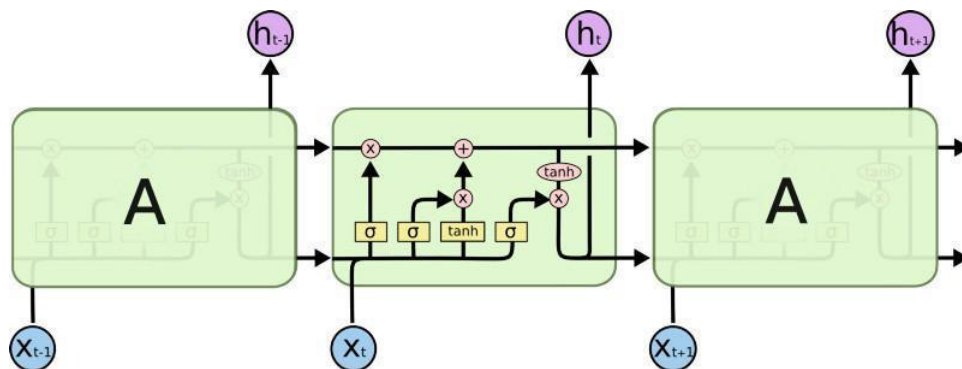


Figura 12 Arquitectura de la unidad LSTM

Capítulo III: Estado del arte

Actualmente en los sistemas de vigilancia se están aplicando diversas soluciones apoyadas con las diferentes Tecnologías de Información para obtener mejores resultados en su monitoreo; buscando la automatización de tareas como la detección y reconocimiento de actividades humanas, principalmente relacionadas con situaciones de inseguridad ciudadana con presencia de violencia y que sean totalmente independientes de cualquier intervención humana. Para un mejor entendimiento de las tecnologías que nos podrían ayudar a abordar este problema, se realizó una búsqueda especializada de artículos científicos, seleccionando 17 papers para validar y resolver el problema, la mayoría de ellos enfocados en el campo de la Visión Artificial, permitiéndonos así revisar las técnicas que son actualmente usadas en este campo. Los artículos seleccionados brindan información en distintos escenarios, en la figura se presenta el resumen de los artículos presentados.

Tabla 3 Resumen del Estado del Arte

Tipo	# Paper	Título	Autor	País	Año	Fuente
Modelos para la detección de eventos	1	Smart surveillance based on video summarization	Sinnu Susan Thomas, Sumana Gupta, Venkatesh K. Subramanian	India	2017	2017 IEEE Region 10 Symposium (TENSYP)
	2	Protest Activity Detection and Perceived Violence Estimation from Social Media Images	Donghyeon Won, Zachary C. Steinert-Threlkeld and Jungseock Joo	USA	2017	MM '17 Proceedings of the 25th ACM international conference on Multimedia
	3	Learning to detect violent videos using convolutional long short-term memory	Swathikiran Sudhakaran and Oswald Lanz	Italia	2017	2017 14th IEEE International Conference on Advanced Video and Signal Based

						Surveillance (AVSS)
Métodos de detección en secuencias de imágenes	4	Intelligent video surveillance for real-time detection of suicide attempts	WassimBouachira, RafikGouiaab, BoLib, RitaNoumeirb	Canadá	2018	Journal-Elseiver (Pattern Recognition Letters)
	5	Human Weapon-Activity Recognition in Surveillance Videos Using Structural-RNN	Praneeth Susarla, Utkarsh Agrawal, Dinesh Babu Jayagopi	USA	2018	MedPRAI '18 Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence
	6	Moving object detection in videos using principal component pursuit and convolutional neural networks	Enrique D. Tejada and Enrique D. Tejada	Canadá	2018	2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)
	7	Face recognition in real-world surveillance videos with deep learning method	Ya Wang, Tianlong Bao, Chunhui Ding, Ming Zhu	China	2017	2017 2nd International Conference on Image, Vision and Computing (ICIVC)
	8	Abnormal event detection in indoor video using feature coding	Mona Izadi, Zohreh Azimifar, and Gholam-Hossein Jowkar	Iran	2017	2017 Artificial Intelligence and Signal Processing Conference (AISP)
	9	Spatio-temporal feature using optical flow based distribution for violence detection	Amira Ben Mabrouk, Ezzeddine Zagrouba	Tunisia		Journal-Elseiver (Pattern Recognition Letters)
	10	Dynamic detection of abnormalities in video analysis of crowd behavior with DBSCAN and neural networks	Hocine Chebia, Dalila Acheli, Mohamed Kesraoui	Algeria	2016	Journal-ASTESJ(Special issue on Recent Advances in Electrical and Electronics Engineering)

	11	Analysis of Gait Pattern to Recognize the Human Activities	Gupta, J. P., P. Dixit, and V. Bhaskar-Semwal	India	2014	Journal-IJIMAI(Special Issue on Multisensor User Tracking and Analytics to Improve Education and other Application Fields)
Aplicaciones en sistemas de vigilancia	12	A unified smart surveillance system incorporating adaptive foreground extraction and deep learning-based classification	Kahlil Muchtar, Faris Rahman, Muhammad Rizky Munggaran, Alvin Prayuda Juniarta Dwiyanoro, Richard Dharmadi, Indra Nugraha	Japan	2019	2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)
	13	HOG and Gabor Filter Based Pedestrian Detection using Convolutional Neural Networks	Fahim Ahmed, Badhon Ahmed Topu, .M. Mohidul Islam	Bangladesh	2019	2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)
	14	Towards Smart Citizen Security Based on Speech Recognition	Jorge Roa, Guillermo Jacob, Leandro Gallino, and Patrick C. K. Hung	Argentina	2018	2018 Congreso Argentino de Ciencias de la Informática y Desarrollos de Investigación (CACIDI)
	15	Semi-supervised dictionary learning via local sparse constraints for violence detection	Tao Zhanga, Wenjing Jiab, Chen Gongc, Jun Suna, Xiaoning Songd	China	2017	Journal-Elsevier (Pattern Recognition Letters)
	16	Automated daily human activity recognition for video surveillance using neural network	Mohanad Babiker , Othman O. khalifa , Kyaw Kyaw Htike , Aisha Hassan , Muhamed Zaharadeen	Malaysia	2017	2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)

	17	Recent advances in video-based human action recognition using deep learning: A review	Di Wu, Nabin Sharma and Michael Blumenstein	USA	2017	2017 International Joint Conference on Neural Networks (IJCNN)
--	----	---	---	-----	------	--

3.1. Modelos para la detección de eventos

En (Thomas et al., 2017) proponen el modelo Human Visual System(HVS) para detectar eventos en los videos de vigilancia, el modelo de atención basado en HVS ayuda a encontrar las regiones sobresalientes de la imagen y esto a su vez lleva a extraer las secuencias de imágenes clave del video, para hacer que el resumen sea más ergonómico y efectivo. Utilizaron los videos de BBC Motion Gallery, UCF101, OVP, Youtube-8M, UrbanTracker, TRECVID y otros archivos de video en línea, siendo 1084 tomas de video de estos conjuntos de datos.

El enfoque del modelo propuesto se divide primero en el marco optimizado, para el resumen perceptivo del video en marcos donde se evidencia la presencia de las propiedades humanas, segundo en el marco de optimización, para la recuperación de video, en lugar de utilizar todo el video para la indexación, la recuperación de video basada en contenido se realiza solo con el cuadro resumido con requisitos de memoria del sistema reducidos. Crearon una base de datos con los cuadros resumidos como un índice para que cuando se recupere un video apropiado con el NN-classifier se pueda identificar las coincidencias con las características extraídas.

En el marco de resumen, los resultados obtenidos en promedio respecto a las tasas de reducción en los videos por (Shih, 2013) y (Pritch et al., 2008) y del enfoque propuesto son 0.282, 0.280 y 0.315 respectivamente, se observa que el enfoque propuesto tiene una mejor tasa de reducción, de la misma manera en la tasa de información donde (Pritch et al., 2008) obtuvieron 0.365 y el enfoque propuesto 0.080. La tasa de información IR transmite la cantidad de información en el

resumen que mide la eficiencia del marco resumido y tasa de reducción RR se denomina proporción de cuadros en resumen al total de cuadros en el video. En el marco de recuperación se comparan los valores de precisión de la técnica de recuperación mediante la visualización basada en diagrama de caja, observando que los valores de precisión del 50% se encuentran en el tercer y cuarto cuartil siendo el lado más alto de los valores de precisión del enfoque propuesto, concluyendo que este enfoque funciona mejor que los métodos de (Chun et al., 2008) y (Y. H. Lai & Yang, 2015).

En (Won et al., 2017) crítica a los trabajos (Chen et al., 2011) y (De Souza et al., 2010) que proponen clasificar automáticamente las actividades violentas en imágenes o videos, pero solo centran en las agresiones físicas, y no en la violencia percibida en actividades de protesta, motines, manifestaciones o incluso reuniones pacíficas. Desarrollaron un modelo visual novedoso que puede reconocer características sobresalientes de las protestas, especialmente la violencia en las imágenes que se comparten durante las protestas, la forma cómo cambian con el tiempo y el espacio describiendo sus actividades mediante atributos visuales descritos en una escena determinada de manera automática sin la necesidad de la intervención humana, para el conjunto de datos recolectaron 10,000 imágenes que pueden describir una escena de protesta mediante la búsqueda en la web usando un conjunto de palabras clave (por ejemplo, "protesta", "motín", "demostración"), después entrenaron la red neuronal CNN basada en una ResNet de 50 capas con la primera clasificación, luego aplicaron a esta red neuronal muestras aleatorias de su colección de imágenes de Twitter etiquetadas geográficamente sin filtrarlas por palabras clave, y obtuvieron un conjunto de imágenes cuyas puntuaciones de predicción estaban por encima de un umbral, el umbral óptimo se seleccionó empíricamente de un conjunto de imágenes etiquetadas a mano, este conjunto contiene muchos ejemplos negativos, como flash mobs; estos dos conjuntos de imágenes

se fusionaron y se proporcionaron a los anotadores de Amazon Mechanical Turk, quienes etiquetaron la presencia o ausencia de un manifestante en cada imagen.

Se midió la precisión del modelo al estimar la violencia percibida y los sentimientos de las imágenes en las dimensiones emocionales, teniendo los coeficientes de correlación lineal de Pearson para la violencia percibida de 0.900 y para la predicción de sentimientos como el enojo, miedo, tristeza y feliz de 0.753, 0.626, 0.340 y 0.382 respectivamente y los coeficientes de determinación (r^2) para la violencia percibida de 0.809 y para la predicción de sentimientos como el enojo, miedo, tristeza y feliz de 0.566, 0.392, 0.116 y 0.146 respectivamente.

En (Sudhakaran & Lanz, 2017) se critica a (Dong et al., 2016) quienes para que detecten los cambios espaciales en un área determinada que tiene un video, recurrieron a métodos que implican la adición de más flujos de datos, como las imágenes de flujo óptico, lo que resulta una mayor complejidad computacional. Desarrollaron un modelo de red neuronal profunda entrenable de extremo a extremo, esta red neuronal consta de capas convolucionales, normalización y agrupación (azul) y el Hidden State, al ser una convLSTM (red neuronal convolucional con la memoria convolucional a largo plazo), es capaz de capturar características espacio-temporales localizadas que permiten el análisis del movimiento local que tiene lugar en el video, generando una mejor representación con un menor número de parámetros, para detectar la presencia de violencia en un video.

Para evaluar la efectividad de la precisión de clasificación si existe violencia en los videos, utilizaron tres conjuntos de datos públicos estándar: Hockey Fight Dataset, Movies Dataset y Violent-Flows Crowd Violence Dataset, extrajeron de cada video, un número N de fotogramas equidistantes en el tiempo que se redimensionan a una dimensión de 256×256 para el entrenamiento, evitando los cálculos redundantes involucrados en el procesamiento de todas las

imágenes que forman un vídeo, ya que las imágenes adyacentes contienen información superpuesta, la cantidad de cuadros seleccionados se basa en la duración promedio de los videos presentes en cada conjunto de datos, dado que la cantidad de videos presentes en los conjuntos de datos es limitada, utilizan técnicas de aumento de datos como el recorte aleatorio y el giro horizontal de las imágenes, durante la etapa de entrenamiento en cada iteración antes de aplicarla en la red, teniendo en cuenta que se sigue la misma técnica de aumento para todos los cuadros presentes en un video, la red se ejecuta para 7500 iteraciones durante la etapa de entrenamiento, en la etapa de evaluación, los cuadros de video se redimensionan a 224×224 y se aplican a la red para clasificarlos como violentos o no violentos.

En los resultados se puede observar los valores de precisión de clasificación obtenidos para los diversos conjuntos de datos considerados en el estudio y se compara con 10 técnicas de vanguardia según (Sudhakaran & Lanz, 2017) (MoSIFT+HIK, ViF, MoSIFT + KDE + Sparse Coding, Deniz et al., Gracia et al., Substantial Derivative, Bilinski et al., MoIWLD, ViF+OVIF, Three streams + LSTM). Donde se puede ver que el método propuesto mejora los resultados de las técnicas existentes en el caso de las series de datos de Hockey Fight Dataset y Movies Dataset con $97.1 \pm 0.55\%$ y $100 \pm 0\%$ respectivamente, sin embargo, en el caso del conjunto de Violent-Flows Dataset, el método propuesto no es capaz de tener los mejores resultados quedando en segundo lugar con $94.57 \pm 2.34\%$, siendo superado por la técnica Bilinski et al. con 96.4% .

También se realiza un estudio comparativo entre el LSTM totalmente conectado tradicional y convLSTM, obteniéndose resultados que muestran que el modelo convLSTM tiene una mejor precisión de clasificación de la existencia de violencia en un video con $97.1 \pm 0.55\%$ en comparación con el LSTM que solo tiene $94.6 \pm 1.19\%$, y además el modelo convLSTM requiere

un menor número de parámetros de 9.6M(9619544) en cambio el modelo tradicional LSTM requiere 77.5M(77520072), evitándose así el sobreajuste excesivo de datos.

En resumen:

- En (Thomas et al., 2017) la detección de eventos se realiza mediante el resumen de video, por lo tanto, esta detección no se realiza en tiempo real, el resumen de video está en función de las propiedades humanas, no necesariamente ligadas a la violencia. En (Won et al., 2017) y (Sudhakaran & Lanz, 2017) para la detección de eventos relacionados con la violencia se puede realizar en tiempo real, utilizan una red neuronal convolucional (CNN), (Won et al., 2017) además de estimar la violencia percibida, puede analizar los sentimientos de imagen en las dimensiones emocionales, (Thomas et al., 2017) y (Sudhakaran & Lanz, 2017) analizan las secuencias de imágenes del video y en (Won et al., 2017) sólo imágenes estáticas.
- En (Thomas et al., 2017) se lleva a cabo una validación experimental para estudiar el rendimiento de la vigilancia inteligente propuesta mediante el resumen de video utilizando seis conjuntos de datos que son los videos de BBC Motion Gallery, UCF101, OVP, Youtube-8M, UrbanTracker, TRECVID, para evaluar el rendimiento del método propuesto por (Sudhakaran & Lanz, 2017) se evalúa en tres conjuntos de datos públicos estándar teniendo a Hockey Fight Dataset, Movies Dataset y Violent-Flows Crowd Violence Dataset.
- En (Sudhakaran & Lanz, 2017) realizaron técnicas de aumento de datos como el recorte aleatorio y el giro horizontal durante la etapa de entrenamiento, (Won et al., 2017) para evaluar el rendimiento construyeron un conjunto de datos novedosos a gran escala, el

Conjunto de datos de imagen de protesta de UCLA, que contiene más de 10 mil imágenes de protesta con sus valores de violencia percibidos anotados manualmente.

3.2. Métodos de detección en videos

En (Bouachir et al., 2018) se menciona que (Lee et al., 2014) presentaron un método que analiza automáticamente las imágenes de profundidad capturadas por una cámara Asus Xtion Pro y detecta comportamientos suicidas, pero solo se consideraron el caso de suspensión parcial sin tratar las dificultades del mundo real, como la oclusión y la invariancia de escala, y para la evaluación del algoritmo solo se utilizaron unas pocas secuencias de video con una duración corta de 3 segundos cada una.

Proponen un sistema de videovigilancia inteligente que utiliza flujos de profundidad proporcionados por una cámara RGB-D, independientemente de las condiciones de iluminación, este sistema está basado en la visión para detección automática del suicidio por ahorcamiento, apoyado en el Algoritmo 1 (Bouachir et al., 2018), para la estimación de parámetros de escala y el Algoritmo 2 (Bouachir et al., 2018) en el reconocimiento de acciones, también proponen el balanceo del conjunto de datos desequilibrado. Realizaron un análisis para la representación y ubicación de las articulaciones humanas por el método de (Shotton et al., 2011), después para garantizar la invariancia de escala, normalizaron cada vector de características F_t mediante un parámetro de escala, para esto implementan el Algoritmo 1 (Bouachir et al., 2018) basado en montones Min-Max para estimar el valor de la mediana en un tiempo constante, para la estimación de los parámetros de escala, en la selección de características y aprendizaje utilizaron la Relevancia Máxima de Redundancia Mínima (mRMR) que es una técnica de filtro popular que es capaz de manejar la redundancia entre la característica seleccionada de manera secuencial respecto a la clase objetivo, el reconocimiento de la actividad en tiempo real se logra mediante el procedimiento resumido por

el Algoritmo 2 (Bouachir et al., 2018) una vez entrenado el clasificador RBF-SVM. Como no existía un conjunto de datos de vídeo público crearon un conjunto de datos de video, donde participaron 21 personas para realizar numerosas simulaciones en una sala cuyas dimensiones son cercanas a las de las cárceles, para el balancear el conjunto de datos de entrenamiento desequilibrado proponen utilizar SMOTE, una técnica de sobre muestreo de minorías sintéticas (Chawla et al., 2002).

Se presentan los resultados obtenidos por 4 clasificadores: Linear Discriminant Analysis, Linear Support Vector Machines, Support Vector Machines y Naive Bayes.

La detección de suicidio basada en la secuencia de vídeo mediante la técnica Leave-One, se observa que el mejor rendimiento se logró utilizando el clasificador L-SVM con el 90% de precisión y el 0% de falsas alarmas y los resultados de la clasificación de fotogramas individuales utilizando el clasificador L-SVM también mediante la técnica Leave-One obtuvo el mejor resultado con el 85% de precisión y el 6% de falsas alarmas, seguido del clasificador RBF-SVM con el 83% de precisión y el 4% de falsas alarmas.

Los resultados obtenidos utilizando un conjunto de datos equilibrado en la detección de suicidio basada en la secuencia de video, la mejor precisión se obtuvo utilizando NB (Naive Bayes) y RBF-SVM con un 100% de precisión con una falsa alarma del 0%, por otro lado, resultados de la clasificación de fotogramas individuales se obtuvo 90% de precisión y el 8% de falsas alarmas, confirmando la efectividad de la estrategia de balanceo de datos para mejorar el rendimiento del sistema.

En (Susarla et al., 2018) se crítica a los sistemas actuales de videovigilancia que necesitan un humano para monitorearlos constantemente como pasa en (P. K. Lai et al., 2016) y (Rota & Thonnat, 2000), y cuando se necesite cambiar su actividad de vigilancia a tiempo real tendría que

hacerse grandes cambios en la arquitectura del sistema, siendo estos cambios costosos. Proponen un sistema de videovigilancia con un reconocedor pre-entrenado, un detector y un rastreador comunes para todos los objetos de vigilancia, para tratar todos ellos simultáneamente siguiendo un enfoque de Interacción Humano-Objeto HOI, el sistema utilizará el método de redes neuronales recurrentes estructurales (SRNN) para aprender automáticamente tales interacciones, hacen uso de datos RGB en lugar de RGB-D.

Considerando la vigilancia en interiores crearon un conjunto de datos de vigilancia RGB (IISD) similar a CAD-120, IISD contiene 55 videos en total formados de 5 (3 mujeres, 2 hombres) sujetos diferentes, las especificaciones detalladas de los datos de IISD:

Estadísticas de IISD:

Total no. de videos - 55	Total no. de sujetos (3 mujeres 2 hombres) - 05
--------------------------	---

Total no. de actividades de alto nivel - 06	Total no. de armas involucradas - 02
---	--------------------------------------

Total no. de sub-actividades etiquetas - 10	Total no. de etiquetas de precio de armas - 08
---	--

Etiquetas de subactividad: caminar, alcanzar, levantar, sostener, sostener y caminar, disparar y caminar, disparar, colocar, salir, estacionario

Actividades de alto nivel de IISD: manejo del arma, patrullando con el arma, soltando el arma, caminando con el arma, disparando con el arma, caminando normal

Armas: Ak-47, Revolver

Etiquetas de armas asequibles: estacionarias, móviles, alcanzables, elevables, sostenibles, disparando, colocables, permutables

Representaron el problema de la vigilancia como un gráfico s-t en el conjunto de datos IISD, por lo tanto, en este gráfico s-t habrá tres nodos (humano, AK-47 y revólver) de dos tipos diferentes (humano, arma), luego modelan una arquitectura de SRNN para tratar el gráfico s-t anteriormente representado, a través del subjectwise folding entrenaron la red SRNN para 1000 iteraciones, con la tasa de aprendizaje de 0,001 en forma de validación cruzada, también realizaron pruebas del conjunto de datos en cada pliegue de validación cruzada después de cada 10 iteraciones del entrenamiento del modelo correspondiente.

En los resultados se puede observar las precisiones de las anotaciones del cuadro delimitador predichas sobre su verdad fundamental utilizando la medida de IOU(Intersección sobre Unión) para un video de 55 segundos en el conjunto de datos de IISD, se tiene una precisión de 94.597% en solo detección, 94.486% en detección con seguimiento de KCF sobre 5 cuadros, 92.249% en detección con seguimiento de KCF sobre 10 cuadros, 93.106% en detección con seguimiento de KCF sobre 15 cuadros, 94.436% en detección con seguimiento MIL sobre 5 cuadros, 92.765% en detección con seguimiento MIL sobre 10 cuadros, 93.699% en detección con seguimiento MIL sobre 15 cuadros.

En (Tejada & Rodriguez, 2017) crítica que existen otros modelos basados en CNN para la detección y clasificación de objetos como lo proponen en (Zhou et al., 2013) y (Kang et al., 2016), teniendo el inconveniente de que analizan los videos en modo batch, también se menciona que hay técnicas de preprocesamiento que mejoran el rendimiento de clasificación, como el método propuesto en (Lawrence et al., 1997), pero según el conocimiento de (Tejada & Rodriguez, 2017), no se han informado técnicas de preprocesamiento para el caso en el que el objetivo es clasificar los objetos en movimiento en secuencias de video.

Proponen utilizar el método Robust PCA (RPCA, también conocido como Principal Component Pursuit, PCP), como un paso de procesamiento previo de modelado de fondo de video, antes de usar el modelo de red neuronal Faster R-CNN, mejorando el rendimiento general de detección y clasificación de los objetos en movimiento.

Seleccionaron el conjunto de datos CDNet2014 para las pruebas, en la clasificación de objetos en movimiento en videos, se realiza primero un modelado de fondo de video basado en el método RPCA / PCP, con un enfoque PCP incremental en el algoritmo que se adapta a los cambios de fondo de la imágenes que forman el vídeo, como los cambios repentinos de iluminación, una vez

realizado el preprocesamiento por medio del algoritmo, se envían las imágenes con menos regiones de interés a un CNN pre-entrenado, específicamente el modelo Faster-RCNN, reduciendo el costo computacional y de clasificación, gracias al preprocesamiento previo de PCP, finalmente la red neuronal devuelve los cuadros delimitadores en las imágenes con los objetos detectados.

En los resultados obtenidos se logra demostrar que en 5 de los 7 conjuntos de datos (backdoor, busStation, pedestrians, PETS2006, skating), existe una mejora que va del 3,7% al 97,2% en el rendimiento de la clasificación de las imágenes con preprocesamiento respecto a las imágenes originales, gracias a la segmentación de los objetos en movimiento por medio del algoritmo PCP. En los otros dos conjuntos de datos (cubicle, highway) no existe mejora alguna.

En (Ya et al., 2017) se critica a los métodos de reconocimiento facial tradicionales que se basan en la representación de características dada por descriptores hechos a mano como LBP (Ahonen et al., 2004), Gabor (W. Zhang et al., 2005), eigenfaces (Turk & Pentland, 1991), que sólo se desempeñan bien en entornos controlados, pero no en el mundo real debido a las grandes variaciones en la iluminación, la postura y la baja resolución. Proponen un método para el reconocimiento facial en vídeos de vigilancia del mundo real utilizando el modelo de cara VGG, donde en la primera parte, construyeron un conjunto de datos mediante la recopilación y el etiquetado automático de los datos, en la segunda parte, ajustan el modelo de reconocimiento facial VGG mediante el conjunto de datos creados.

Crearon un nuevo conjunto de datos mediante la recopilación y el etiquetado automático de los datos de un video de vigilancia del mundo real, el conjunto de datos se establece en cuatro etapas: en la etapa 1 generaron datos en bruto mediante la detección de rostro y seguimiento, la etapa 2 se realiza la purificación dentro de cada clase mediante la agrupación de gráficos por característica VGG, la etapa 3 también se realiza la purificación pero entre clases, con la medida de similitud,

finalmente en la etapa 4 se realiza el filtrado según cantidad eliminando las clases identidad para las cuales no hay suficientes imágenes, con el nuevo conjunto de datos ya establecidos, realizan un ajuste fino a el modelo de cara VGG, este modelo comprende ocho capas convolucionales y tres capas totalmente conectadas, a cada una de ellas les sigue una o más no linealidades, como ReLU y max-pooling, los pesos del modelo de cara VGG pre-entrenado se ajustan al continuar la propagación hacia atrás con el nuevo conjunto de datos.

En el conjunto de datos de prueba, el modelo de rostro VGG después de un primer ajuste fino con una escala de 140 en las entidades clases, se alcanza una tasa de reconocimiento del 91,4%; en un segundo ajuste fino con una escala de 240 en las entidades clases, se alcanza una tasa de reconocimiento del 92,1%; en todos los ajustes finos realizados anteriormente se supera al modelo de cara VGG sin ajuste fino, cuya tasa de reconocimiento era del 83,6%.

En (Izadi et al., 2018) se menciona que la identificación de eventos anormales de videos de acuerdo al tipo de aprendizaje se podría clasificar en enfoques basados en el seguimiento y no basados en el seguimiento, donde los enfoques basados en el seguimiento son adecuados para escenas o videos con una pequeña cantidad de objetos como se realiza en (Morris & Trivedi, 2011), sin embargo, no son aplicables para la detección de patrones anormales en escenas complejas o abarrotadas. Proponen un sistema para la detección de eventos anormales de videos de vigilancia, centrado en un algoritmo de visión de computadora no supervisado en escenas dinámicas, incluyendo escenas de áreas abarrotadas o con poca gente, este algoritmo se basa en la codificación dispersa original sin ninguna optimización y proporciona vectores de reconstrucción y normalidad basados en estos vectores, también utilizan un diccionario (conjunto de bases) para adaptarse a datos específicos para la detección de eventos anormales, buscando aumentar la precisión de los métodos disponibles mediante la extracción de características robustas.

En este trabajo, utilizaron dos conjuntos de datos de video disponibles de las cámaras de vigilancia de entrada y salida de las estaciones de metro de (Adam et al., 2008). El proceso para la detección de eventos anormales en los videos, comienza con la detección de los puntos de interés en las secuencias de video a través del algoritmo de (Dollar et al., 2005), luego los cubos espacio-temporales obtenidos con el algoritmo correspondientes a cada punto de interés se descomponen y con la combinación de los descriptores HOG y MBH se extraen los vectores de características. Se tiene un diccionario inicial de la imagen (que se aprende en base a los primeros cinco minutos del video) y los vectores de características obtenidos se utilizan para actualizar el diccionario. En el último paso se calcula el escaso costo de reconstrucción (SRC), que de acuerdo a la función definida se basa en un valor de umbral que clasifica un evento como anormal o normal.

En los resultados se observa la reducción del número de falsas alarmas en la detección inusual de eventos en el video de vigilancia de la entrada del metro. WD: dirección incorrecta; NP: sin pago; LT: merodeando; II: interacciones irregulares; MISC: incluyendo parada repentina, corriendo rápido, donde el número de falsas alarmas del método propuesto es 1 y respecto a los métodos de (J. Kim & Grauman, 2010) y (Zhao et al., 2011) su número de falsas alarmas son 3 y 2 respectivamente. También sucede de la misma manera en la detección inusual de eventos en el video de vigilancia de salida de metro. WD: dirección incorrecta; LT: merodeando; MISC: incluyendo parada repentina, corriendo rápido donde el número de falsas alarmas del método propuesto es 3 y respecto a los métodos de (J. Kim & Grauman, 2010) y (Zhao et al., 2011) su número de falsas alarmas son 6 y 5 respectivamente.

En (Ben Mabrouk & Zagrouba, 2017) se critica a los métodos locales de las investigaciones [72][73] [94][95][96][97][98] que extraen puntos de interés o características espacio-temporales para describir la acción humana en un video, sin embargo es posible que no produzcan

información significativa sobre la acción cuando hay demasiado movimiento, también se critica a los métodos globales de las investigaciones [99][100][101][102] que extraen características globales como el flujo óptico para describir el movimiento en todo el cuadro, pero debido a que son sensibles al ruido y a los fondos saturados llevando a obtener información irrelevante sobre la acción. Proponen un método para la detección de violencia, donde teniendo un vídeo de entrada se detecta primero un conjunto de puntos STIP para determinar los marcos de interés que se utilizan para extraer nuestro descriptor, luego, extrajeron la función DiMOLIF basada en puntos STIP e información de flujo óptico, finalmente las características extraídas se utilizan como una entrada para el clasificador SVM.

En el proceso de extracción de características del método propuesto, primero para cada marco de interés se dividieron en 4×4 bloques no superpuestos, cada bloque contiene un conjunto de puntos STIP, donde solo interesan los bloques que contienen un número suficiente de puntos STIP, por lo tanto aplican un algoritmo de filtrado para descartar bloques irrelevantes que tienen un número de puntos STIP por debajo de un umbral N, en el siguiente paso modelan los cambios de movimiento en cada bloque a lo largo del tiempo formando un cubo temporal de espacio (STC) que representa cada bloque mediante la recopilación de T cuadros continuos, cada STC se construye alrededor del centroide de todos los puntos STIP incluidos en el bloque, luego, se calcula el vector de flujo óptico para estimar el movimiento de los píxeles entre dos cuadros consecutivos, después se muestra la distribución para obtener una curva bidimensional, la distribución se cuantifica para obtener un histograma de 8 bandejas que se concatenan y normaliza, finalmente obteniendo la función de extracción de características DiMOLIF de la dimensión 128 ($4 \times 4 \times 8$). Usaron el conjunto de datos Hockey Fight y el conjunto de datos de Flujos para probar la efectividad de la

detección de violencia del método DiMOLIF propuesto, tanto en escenas abarrotadas como con poca gente.

En el conjunto de datos de Hockey Fight, los resultados demuestran que el método propuesto es más eficiente para detectar la violencia en una escena con poca gente con una precisión de la predicción media (ACC) $\pm$ la desviación estándar (SD) de $88.6 \pm 1.2\%$, superando a los demás métodos ViF, OViF, ViF+OViF que obtuvieron una precisión de la predicción media (ACC) $\pm$ la desviación estándar (SD) de $81.60 \pm 0.22\%$, $84.20 \pm 3.33\%$, $86.30 \pm 1.57\%$ respectivamente.

En cambio en el conjunto de datos de Flujos violentos que está especialmente diseñado para evaluar la violencia en una escena concurrida, el método propuesto obtiene una precisión de la predicción media (ACC) $\pm$ la desviación estándar (SD) de $85.83 \pm 4.26\%$, que es superior en comparación con los métodos de ViF y OViF que obtuvieron una precisión de la predicción media (ACC) $\pm$ la desviación estándar (SD) de $81.2 \pm 1.79\%$, $76.8 \pm 3.9\%$ respectivamente, sin embargo es superado por el método ViF+OViF que obtuvo $86.00 \pm 1.41\%$ en la predicción media (ACC) $\pm$ la desviación estándar (SD).

En (Chebi et al., 2016) se critica los enfoques existentes (Mahadevan et al., 2010), (Pathan et al., 2010) y (Ko, 2008) que solo detectan anomalías en un grupo o una sola persona, sin poder hacer la detección de anomalías en ambos casos (un grupo o una sola persona). Se propone un nuevo enfoque para la detección de anomalías dinámicas de escenas muy densas que miden la velocidad de los individuos y de todo el grupo, las diversas anomalías se detectan cambiando dinámicamente entre dos enfoques: una red neuronal artificial (ANN) para la gestión de anomalías grupales de personas y una agrupación espacial de aplicaciones con ruido (DBSCAN) basada en densidad en el caso de las entidades, para una mayor robustez y eficacia, introducen dos rutinas que sirven para eliminar las sombras y el manejo de las oclusiones.

Realizaron una arquitectura detallada en varias secciones; se tiene la sección de Extracción de vectores de movimiento, donde las actividades de multitud investigadas se caracterizan por el movimiento de personas, también está la sección Eliminación de la sombra, donde se hace la reducción de los tonos que se traduce como un simple paso del espacio colorimétrico RGB hacia el espacio HSV, otra sección que tiene la arquitectura es la sección Oclusión de gestión, la cual desarrolla una metodología de segmentación híbrida para dividir burbujas ocluidas utilizando un enfoque basado en histogramas para la oclusión horizontal y un enfoque basado en elipse para la oclusión vertical. También Existen dos secciones de agrupación de vectores de movimiento, la primera usa DBSCAN para agrupar los vectores de movimiento en diversos grupos de puntos de datos que tienen coordenadas similares, magnitudes similares y orientación similar, donde cada grupo de vectores de movimiento representa un patrón de movimiento y la segunda que utiliza una red neuronal para clasificar el comportamiento de la multitud, finalmente es la sección de detección de eventos donde la estrategia de detección se representa como una forma dinámica para detectar el comportamiento de la multitud mediante dos enfoques (DBSCAN y ANN) ya sea el caso de la entidad y el grupo de personas. Los videos se recopilaron principalmente del conjunto de datos UMN, el conjunto de datos de video BEHAVE y el conjunto de datos PETS2009, para la evaluación del rendimiento se adoptan en experimentos de detección de comportamiento de cuadros anómalos.

Entre los resultados más ilustrativos se observa que para el caso de eliminación de sombra, el enfoque propuesto es favorable, ya que presenta una contribución positiva para la detección de las razones del movimiento en un entorno complejo, mostrando que el sistema tiene una buena robustez con una precisión de más de 0,75.

En (Prakash Gupta et al., 2014) proponen un método para el sistema de reconocimiento automático de la actividad humana a través de la marcha (identificar una actividad por la forma en que caminan) sin intervención humana, este sistema se basa en la extracción de primer plano, el seguimiento de personas, la extracción de características y el reconocimiento. El método propuesto que utilizara el Sistema para el reconocimiento de la actividad humana identifica cuatro actividades humanas básicas (caminar, correr, trotar y saltar).

El método propuesto tiene los siguientes pasos principales: extracción en primer plano, seguimiento humano, extracción de características y reconocimiento de actividades. En la extracción en primer plano, se proporciona el video como una entrada al sistema desde la base de datos de actividad y se extraen los marcos de ese video, el modelo paramétrico de humanos se extrae de secuencias de imágenes mediante la detección y el seguimiento de humanos basados en movimiento / textura, después de la extracción de secuencias de imágenes, los resultados muestran las actividades reconocidas como el caminar, correr, trotar y saltar, y finalmente se prueba el rendimiento del método de forma experimental utilizando los conjuntos de datos en ambientes interiores y exteriores.

Para evaluar el enfoque propuesto de reconocimiento de actividad humana, utilizaron dos conjuntos de datos; el conjunto de datos de Acciones humanas de KTH("caminar", "trotar", "correr", "boxear", "agitar las manos" y "aplaudir con las manos") y conjunto de datos de Acciones de Weizmann("correr", "caminar", "trotar", "saltar", "saltar adelante en dos piernas", "saltar en el lugar" "en dos piernas", "galopando lateralmente", "onda dos manos", "onda una mano" o "curva")

El enfoque propuesto por los autores tiene una precisión hasta el 95.01% en el correcto reconocimiento de las actividades humanas en el conjunto de datos KTH ("caminar", "trotar",

"correr") y el 91.36% de precisión en el correcto reconocimiento de las actividades humanas en el conjunto de datos Weizmann("caminar", "trotar", "saltar").

Observando los resultados de precisión obtenidos (95% y el 91%), se muestra que el método propuesto supera a otros métodos propuestos por (R. Zhang et al., 2007), (Vega & Sarkar, 2003) y (Noorit et al., 2010) que lograron un 61%, 90% y 93% de precisión respectivamente.

En resumen:

- En (Chebi et al., 2016), (Babiker et al., 2018), (Susarla et al., 2018), (Ya et al., 2017) y (Tejada & Rodriguez, 2017) en su propuesta de solución utilizaron las redes neuronales como principal componente, de donde tenemos a (Ya et al., 2017) y (Tejada & Rodriguez, 2017) que utilizaron el tipo de red neuronal convolucional, (Ya et al., 2017) utilizó en específico el modelo VGG16 para el reconocimiento facial en video de vigilancia y (Tejada & Rodriguez, 2017) el modelo Faster R-CNN para la detección de objetos en movimiento junto al algoritmo Robust PCA (RPCA, también conocido como Principal Component Pursuit, PCP); (Susarla et al., 2018) utilizó el tipo de red neuronal estructural recurrente para el reconocimiento de hombres armados, finalmente (Chebi et al., 2016) y (Babiker et al., 2018) utilizaron solo las redes neuronales artificiales tal cual, (Chebi et al., 2016) para la detección dinámica de anomalías del comportamiento de un grupo de personas intercambiando con DBSCAN si se trata de una persona y (Babiker et al., 2018) para la clasificar el tipo de actividad humana que se está realizando mediante la videovigilancia.
- Tenemos los trabajos (Bouachir et al., 2018) (Izadi et al., 2018) y (Ben Mabrouk & Zagrouba, 2017) no utilizaron redes neuronales artificiales, si no en cambio (Izadi et al., 2018) y (Ben Mabrouk & Zagrouba, 2017) se basaron en el flujo óptico mediante STIP, (Izadi et al., 2018) para la detección de eventos anormales en interiores y (Ben Mabrouk &

Zagrouba, 2017) para la detección de violencia. Podemos observar que en (Bouachir et al., 2018) donde utilizaron un algoritmo propuesto se basa en la explotación de las posiciones de las articulaciones del cuerpo para modelar el comportamiento suicida, está relacionada con la detección de violencia al igual que (Ben Mabrouk & Zagrouba, 2017) a diferencia de (Chebi et al., 2016), (Babiker et al., 2018), (Susarla et al., 2018), (Ya et al., 2017), (Tejada & Rodriguez, 2017) y (Izadi et al., 2018), en (Bouachir et al., 2018) Sude detecta la violencia autoinfligida y en (Ben Mabrouk & Zagrouba, 2017) la violencia interpersonal.

- En (Susarla et al., 2018), (Ya et al., 2017) y (Bouachir et al., 2018) crearon su propio conjunto de datos de video, en el cual (Susarla et al., 2018) crearon un conjunto de datos de vigilancia RGB (IISD) similar a CAD-120, En (Ya et al., 2017) construyeron un conjunto de datos de manera novedosa mediante un proceso de detección de rostros, seguimiento y agrupación de gráficos y finalmente tenemos a (Bouachir et al., 2018) dado que no existía un conjunto de datos de video público para evaluar el sistema propuesto, se tuvo que crear uno basado en escenarios de suicidio.

3.3. Aplicaciones en Sistemas de Vigilancia

En (Kahlil et al., 2019) critica el trabajo propuesto por (C. Kim et al., 2018) que introduce un híbrido utiliza GMM para una resta de fondo y encontrar la región de interés (ROI), pero el método GMM es propenso al ruido y muy sensible a los cambios de iluminación, además de esta condición, también el procesamiento de imágenes del trabajo propuesto por (C. Kim et al., 2018) es por píxeles, siendo no adecuado para la detección en tiempo real. Se propone un Sistema basado en una nueva técnica unificada (el buscador de región de interés (ROI) y YOLO que es un detector de profundidad), para detectar un objeto en movimiento a partir de los videos de vigilancia basados en la CPU (unidades de procesamiento central).

Analizaron dos conjuntos de datos públicos para evaluar el sistema propuesto, el primer conjunto de datos es CDNET2014, que presenta varias escenas desafiantes que incluyen mal tiempo, baja velocidad de fotogramas, sombra, línea de base, por nombrar algunas, seleccionaron dos videos representativos, que son autopistas y peatones, respectivamente, el segundo conjunto de datos es UCF-Sports que ofrece varias acciones deportivas, por ejemplo, personas que montan a caballo, caminan, etc. Para el diseño de algoritmos, eligieron un enfoque basado en texturas de bloques determinando las áreas de los objetos en movimiento a través del buscador de ROI, que identifica las regiones de interés por ende se minimiza la región analizar. Una vez identificada la región de interés para la detección de profundidad aplicamos el detector YOLO, quien realiza la predicción de la coordenada de los cuadros delimitadores con su propia puntuación de confianza, que se calcula al agregar las probabilidades, por lo tanto, un número menor de celdas de la cuadrícula de un cuadro entrante, será una detección más rápida, finalmente para mejorar la eficiencia de reconocimiento se hace uso de las unidades modernas de procesamiento de gráficos (GPU).

En un total de 1696 frames evaluaron la velocidad de procesamiento de las secuencias de imágenes del conjunto de datos de carreteras CDNET2014 que tiene un tamaño de frame de 320x240, donde el Modelo Yolov3 con 80 clases de objeto obtuvo de 14,74 frames/sec y el modelo propuesto para vehículos con 24 clases de vehículo obtuvo 22,02 frames/sec y con un tamaño de frame propuesto por el buscador de ROI en ambas evaluaciones se observa mejoras significativas en la velocidad de procesamiento de las secuencias de imágenes tanto en el Modelo Yolov3 con 16,62 frames/sec como el modelo propuesto para vehículos con 24,57 frames/sec.

En (Ahmed et al., 2019) propusieron un sistema de detección peatonal basado en el filtro gaussiano para eliminar el ruido gaussiano, HOGG (Filtro de Gabor y Histograma de Gradientes Orientado (HOG)) para la extracción de características, obteniendo mejores resultados que solo HOG o solo

el filtro Gabor, y finalmente las Redes Neuronales Convolucionales (CNN) para la detección de peatones.

Utilizaron tres conjuntos de datos diferentes, el Conjunto de datos de INRIA, DAIMLER Mono Pedestrian Classification Benchmark Dataset que se utilizan para fines de entrenamiento y prueba, y PennFudanPed se usa solo para pruebas. Realizaron una arquitectura, donde primeramente hicieron un preprocesamiento de las imágenes utilizando el filtro gaussiano reduciendo el ruido de la imágenes ubicando los bordes, preservando todos los detalles útiles de la imagen, después utilizaron el método de HOG para la extracción de características, desarrollaron 3 modelos diferentes de redes neuronales convolucionales para encontrar un mejor modelo para la detección de peatones, utilizando las características anteriormente extraídas en los mapas de características en las capas convolucionales de cada uno de los modelos. El primer modelo utilizó tres capas convolucionales y tres capas ocultas, la forma de entrada es 64x64, la entrada se pasa a la capa convolucional que tiene 32 mapas de características con un tamaño de kernel de 3x3 filtros que reducen la dimensión de la imagen a 62x62, utilizaron la unidad lineal rectificadora (ReLU) como función de activación. En el segundo modelo cambiaron a los mapas de características en la capa de convolución, utilizaron tres capas convolucionales y tres capas ocultas como antes, pero en las capas convolucionales, tomaron los mapas de características de tamaño 32, 64, 128 respectivamente. Y en el Modelo 3 utilizaron cuatro capas convolucionales y seis capas ocultas, las capas son similares al modelo 1, excepto en el número de capas.

En los principales resultados se observa que en el conjunto de datos INRIA contiene 5203 imágenes en total se evaluaron los tres modelos donde el modelo 2 siempre muestra la mayor precisión. obteniendo una precisión del 99.81% con una pérdida de valor del 4.34%, es mejor que los métodos propuestos por (Fukui et al., 2015), (Li et al., 2018) y (Liu et al., 2016). En Daimler

Mono Pedestrian Classification Benchmark Dataset que contiene 22404 imágenes trabajaron con 7000 imágenes, evaluaron los tres modelos y se observó que el modelo 2 siempre muestra la mayor precisión. Obtuvieron un 86.8% de precisión, lo que es mejor los métodos propuestos en (Dominguez-Sanchez et al., 2017) y (Nilsson et al., 2014) pero (Stamos et al., 2015) logró un mejor resultado que el método propuesto para el conjunto de datos Daimler.

En (Roa et al., 2018) critica que en el reconocimiento de voz (Da Silva et al., 2016) describen la implementación de un sistema de reconocimiento de voz aislado independiente del orador en tiempo real utilizando modelos ocultos de Markov, con respecto a las características de los audios los enfoques de (Giannakopoulos et al., 2006), (Schedi et al., 2015) y (Bautista-Durán et al., 2017) solo consideran las señales de audio y dejan de lado la semántica y el significado estadístico del discurso que se analiza, para el análisis de texto los enfoques de los autores (Schmidt & Wiegand, 2017), (Davidson et al., 2017), (Munezero et al., 2014), (Kiktova-Vozarikova et al., 2015) y (Farkhadov et al., 2018) no están enfocados en detectar situaciones de inseguridad y no son compatibles con el idioma español.

Proponen un componente de software prototipo para la detección automática de situaciones de inseguridad basadas en el reconocimiento de voz, el procesamiento del lenguaje natural y el reconocimiento de patrones, este componente proporciona un servicio de vigilancia automática que permite detectar en tiempo real las situaciones de inseguridad, generar alertas de seguridad para ayudar a mitigar dichas situaciones e informar y/o registrar los eventos ocurridos.

Obtuvieron los requisitos funcionales y no funcionales para el desarrollo del componente de software, realizan la arquitectura de tubería, donde identificaron 3 componentes principales del componente de software; primero tenemos al reconocedor de voz, que recibe como entrada el audio obtenido del entorno y su salida son las frases textuales reconocidas, luego tenemos el

preprocesador, donde su entrada son las frases reconocidas en la etapa anterior y su salida es el texto conceptualizado y finalmente el detector de situaciones de inseguridad que contiene la estrategia y las reglas para la detección de patrones, se recibe como entrada el texto conceptualizado y devuelve un conjunto de acciones predefinidas para realizar si es o no una situación de inseguridad ciudadana. Los datos de entrada para las pruebas provienen de las películas argentinas que contienen situaciones de violencia en formato digital, de cada película se obtuvieron subtítulos, para su validación, y también de las grabaciones de cámaras de seguridad en las que se produjo alguna situación de inseguridad de YouTube. la principal diferencia entre ambos tipos de fuentes de datos está en la calidad del sonido con el que se grabaron los videos.

En los resultados en casos de pruebas para entornos simulados, se obtuvieron coincidencias entre EAA(Número de activaciones de alarma esperadas antes de la prueba) y OAA(Número de activaciones de alarma obtenidas después de la prueba) para los casos de prueba 2, 3 y 5 que son videos con intervalos de tiempo de 150, 120, 130 segundos respectivamente, sin embargo, para los casos de prueba 1 y 4 que también son videos con intervalos de tiempo de 42, 135 segundos respectivamente, no hubo coincidencia alguna entre EAA y OAA, para los casos de pruebas para entornos controlados, se obtuvieron coincidencias entre EAA y OAA para todos los casos de prueba 1, 2, 3 y 5 que son los mismos videos de los casos de pruebas para entornos simulados. En los casos de pruebas para un ambiente real no se obtuvieron coincidencias en ninguno de los casos entre EAA y OEE, debido a la calidad de entrada de audio, porque los micrófonos de las cámaras de seguridad no son de buena calidad y normalmente se encuentran en lugares que no son muy accesibles, lo que dificulta el reconocimiento de voz.

En (T. Zhang et al., 2018) se criticó la propuesta de (Clarin et al., 2005) porque se centraron en detectar la piel y la sangre en secuencias de vídeo mediante la segmentación en primer plano o la

información del color de la piel, sin embargo, existía una degradación del rendimiento debido a que el color detectado no era lo suficientemente discriminatorio. Proponen un sistema para detectar efectivamente el comportamiento violento, este sistema cuenta un marco de aprendizaje semi-supervisado, que es adecuado para conjuntos de datos muy grandes, un diccionario que aprende de muestras etiquetadas para discriminación, así como un gran número de muestras no etiquetadas y el aprendizaje de datos no etiquetados aumenta aún más su poder representativo, también se propone un esquema de clasificación que integra el modelo disperso modificado.

Desarrollaron un algoritmo de detección de violencia efectivo mediante un algoritmo de aprendizaje de diccionario semi-supervisado que integra el aprendizaje de diccionarios y la capacitación de clasificadores, introdujeron una función objetivo novedosa que contiene términos que representan el error de reconstrucción de los datos etiquetados y no etiquetados a través las restricciones de representación y la incoherencia del coeficiente, este algoritmo alterna entre la codificación dispersa y la actualización del diccionario a medida que las señales de entrada llegan de manera secuencial, una vez que se aprendió el diccionario lo adoptaron para representar una muestra de prueba y realizar una clasificación, para evaluar la precisión de esta clasificación, emplearon la prueba de validación cruzada de 5 veces en cada conjunto de datos, esta evaluación y validación se realizó en tres conjuntos de datos de referencia desafiantes, como son el conjunto de datos de Hockey Fight, el conjunto de datos BEHAVE y el conjunto de datos de Crowd Violence.

Se muestra los resultados de la detección en el conjunto de datos de Hockey Fight, observando que el método propuesto obtuvo una precisión de la predicción media (ACC) $\pm$ la desviación estándar (SD) de $96.5 \pm 1.04\%$, superando a los métodos HOG + BoW, HOF + BoW, HNF + BoW, ViF, MoSIFT + BoW, MoWLD + BoW, AMDN, SRC, MoWLD + Sparse Coding, PSS y SSS que

obtuvieron una precisión de la predicción media (ACC) $\pm$ la desviación estándar (SD) de $88.77 \pm 0.73\%$, $86.07 \pm 0.59\%$, $89.27 \pm 0.79\%$, $90.07 \pm 0.99\%$, $88.8 \pm 0.75\%$, $89.28 \pm 0.93\%$, $89.7 \pm 1.13\%$, $94.2 \pm 1.07\%$, $93.8 \pm 1.08\%$, $95.5 \pm 1.07\%$, $96.1 \pm 1.04\%$ respectivamente.

Los resultados en el conjunto de datos BEHAVE demuestran que el algoritmo SSS es más efectivo para detectar la violencia en una escena de lucha grupal con una precisión de la predicción media (ACC) $\pm$ la desviación estándar (SD) de $89.07 \pm 0.10\%$, superando a los métodos HOG + BoW, HOF + BoW, HNF + BoW, ViF, MoSIFT + BoW, MoWLD + BoW, AMDN, SRC, MoWLD + Sparse Coding, PSS y el método propuesto que obtuvieron una precisión de la predicción media (ACC) $\pm$ la desviación estándar (SD) de $58.97 \pm 0.34\%$, $60.03 \pm 0.28\%$, $58.24 \pm 0.31\%$, $83.62 \pm 0.19\%$, $62.78 \pm 0.23\%$, $81.65 \pm 0.18\%$, $84.22 \pm 0.17\%$, $82.7 \pm 0.14\%$, $85.27 \pm 0.13\%$, $85.15 \pm 0.13\%$, $88.26 \pm 0.11\%$ respectivamente, observando que el método propuesto ocupa el segundo lugar solo superado por SSS.

Los resultados en el conjunto de datos de Crowd Violence demuestran que el método propuesto también es eficaz para detectar la violencia en una escena llena de gente con una precisión de la predicción media (ACC) $\pm$ la desviación estándar (SD) de $92.25 \pm 0.12\%$, superando a los métodos HOG + BoW, HOF + BoW, HNF + BoW, ViF, MoSIFT + BoW, MoWLD + BoW, AMDN, SRC, MoWLD + Sparse Coding, PSS y SSS que obtuvieron una precisión de la predicción media (ACC) $\pm$ la desviación estándar (SD) de $57.98 \pm 0.37\%$, $58.71 \pm 0.12\%$, $57.05 \pm 0.32\%$, $82.13 \pm 0.21\%$, $57.09 \pm 0.37\%$, $88.16 \pm 0.19\%$, $84.72 \pm 0.17\%$, $89.6 \pm 0.18\%$, $89.38 \pm 0.13\%$, $89.5 \pm 0.13\%$, $91.9 \pm 0.12\%$ 0.9357 respectivamente.

En (Babiker et al., 2018) proponen el desarrollo de un sistema inteligente de reconocimiento de la actividad humana, utilizaron una serie de técnicas de procesamiento de imágenes digitales en cada etapa del sistema propuesto, como la sustracción de fondo, la binarización y la operación

morfológica, también construyeron una red neuronal robusta basada en la base de datos de actividades humanas, que extrajeron de las secuencias de imágenes preprocesadas.

Adquirieron la secuencias imágenes de video de una única cámara de video estática, posteriormente se le realiza una serie de operaciones basadas en técnicas de procesamiento de imágenes como la determinación del marco de fondo, resta de fondo, umbral de la imagen utilizando el método Otsu, teniendo una imagen resultante de color blanco y negro(el fondo está representado en color negro, mientras que el objeto estará en color blanco), el filtro de mediana 2D para reducir el ruido y la distorsión en la imagen extraída, la dilatación y la erosión provocada por el proceso de sustracción de fondo realizado por la operación morfológica, finalmente la operación de análisis de manchas para calcular las características de los píxeles blancos, que representan el cuerpo humano, después construyeron una base de datos a partir de la información extraída de la imagen binaria, teniendo un perceptrón multicapa que alimenta la red neuronal utilizada para entrenar el sistema diseñado mediante el mapeo de los datos de entrada y lo conectó con una relación numérica con una salida específica a través de varios nodos de neuronas, una vez que entrenaron el sistema y reconocer su eficiencia, probaron en el sistema una imagen antes de ejecutarlo en el video, ya que lo que funciona con una imagen seguramente también puede funcionar en la secuencia de imágenes que es el contenido de cualquier video. Por último, utilizan la interfaz gráfica de usuario provista en matlab para monitorear todo el proceso, desde la sustracción en segundo plano y el siguiente proceso hasta el seguimiento y reconocimiento de la detección final.

Se muestran los resultados de rendimiento y la tasa de reconocimiento en cada etapa de clasificación en la red neuronal, todo el conjunto de datos de las muestras se divide en muestras de entrenamiento, prueba y validación, el número de muestras en cada una es del 70%, 15%, 15%

respectivamente del número total, que es de 5000 muestras, en la etapa de entrenamiento, se puede ver que 709 de las muestras se clasifican correctamente como actividad de walking, esta correspondencia corresponde al 20,3% del total de 3500 muestras, mientras que 9 muestras se clasifican incorrectamente como actividad de caminar, esta correspondencia corresponde al 0,3% de todas las muestras del conjunto de datos de entrenamiento, además, de 718 muestras, el 98.7% se clasificó correctamente, como actividad de caminar y el 1.3% incorrectamente clasificado, de manera similar pasa con el resto de las actividades (agitar la mano, tendido, caminar, sentarse, boxeo) en las otras etapas.

En resumen:

- En las investigaciones (Kahlil et al., 2019), (Ahmed et al., 2019) y (Babiker et al., 2018) se utilizaron las redes neuronales para la aplicación de los sistemas de vigilancia, (Ahmed et al., 2019) se enfocó en la detección de peatones, (Babiker et al., 2018) en detectar y rastrear el cuerpo humano reconociendo el tipo de movimiento y actividad humana que realiza diariamente (agitar la mano, tendido, caminar, sentarse, boxeo), (Kahlil et al., 2019) solo en la detección de objetos en movimiento, (Kahlil et al., 2019), (Ahmed et al., 2019) y (Babiker et al., 2018) realizaron como paso previo el preprocesamiento de imágenes, (Kahlil et al., 2019) además de realizar técnicas de procesamiento de imágenes, aplica una nueva técnica unificada (el buscador de región de interés (ROI) y YOLO que es un detector de profundidad), mejorando aún más su velocidad de procesamiento de secuencias de imágenes. (T. Zhang et al., 2018) (Roa et al., 2018) si se centraron en la detección de la violencia en los sistemas de vigilancia, (T. Zhang et al., 2018) lo realiza mediante un framework de aprendizaje semi-supervisado para clasificar si un comportamiento es

violento o no, (Roa et al., 2018) a través del audio de las grabaciones de las diferentes fuentes datos dependiendo de la calidad de sonido.

- En (Ahmed et al., 2019) utilizaron principalmente dos conjuntos de datos, el conjunto de datos INRIAPerson y Daimler Mono Peatonal, en (T. Zhang et al., 2018) utilizaron 3 conjuntos de datos, el conjunto de datos de Hockey Fight, el conjunto de datos BEHAVE y el conjunto de datos de Crowd Violence, en (Roa et al., 2018) para su conjunto de datos utiliza el audio de la grabaciones de la diferentes fuentes datos como películas argentinas con situaciones de inseguridad y violencia, y de libros de autores argentinos.

Capítulo IV: Metodología de la investigación

La finalidad de esta investigación consiste en diseñar un modelo de Visión Artificial compuesto por una red neuronal convolucional Xception de Google usada principalmente para la clasificación de imágenes de manera genérica, sin haber intenciones de aplicarlas en la extracción de características en secuencias de imágenes obtenidas en videos de seguridad con situaciones de violencia física interpersonal, también se tiene un módulo de clasificación que utiliza una red neuronal recurrente LSTM con Capas Totalmente Conectadas que permita clasificar la existencia o no de violencia física interpersonal en las secuencias de imágenes de videos.

4.1. Dataset

Debido a las desventajas mencionadas anteriormente de los conjuntos de datos anteriores en la Motivación, decidimos buscar un conjunto de datos con el objetivo de encontrar al menos un conjunto de datos de videos de situaciones de violencia física interpersonal en la vida real, con variedad de personas en raza, edad y género, también variedad en el entorno, terminamos encontrándolo en KAGGLE, que es una comunidad en línea de científicos de datos y estudiantes de máquinas, propiedad de Google LLC, este conjunto de datos fueron recopilados debido a la escasez de los conjuntos grande con una variedad de escenas. El conjunto de datos final actualmente contiene 1000 videos pequeños de violencia y 1000 videos pequeños de no violencia con un total de 2000 videos de una duración máxima de 7 segundos y una duración promedio de 5 segundos. Esto es más grande que el mayor conjunto de datos de la pelea de hockey. Los videos de no violencia fueron recopilados de otros videos grandes de diferentes escenas como personas comiendo, jugando tiro con arco, equitación, ajedrez, tenis, baloncesto, fútbol, voleibol, natación, levantamiento de pesas, gimnasio, billar, escenas de comedia de películas de Charlie Chaplin.

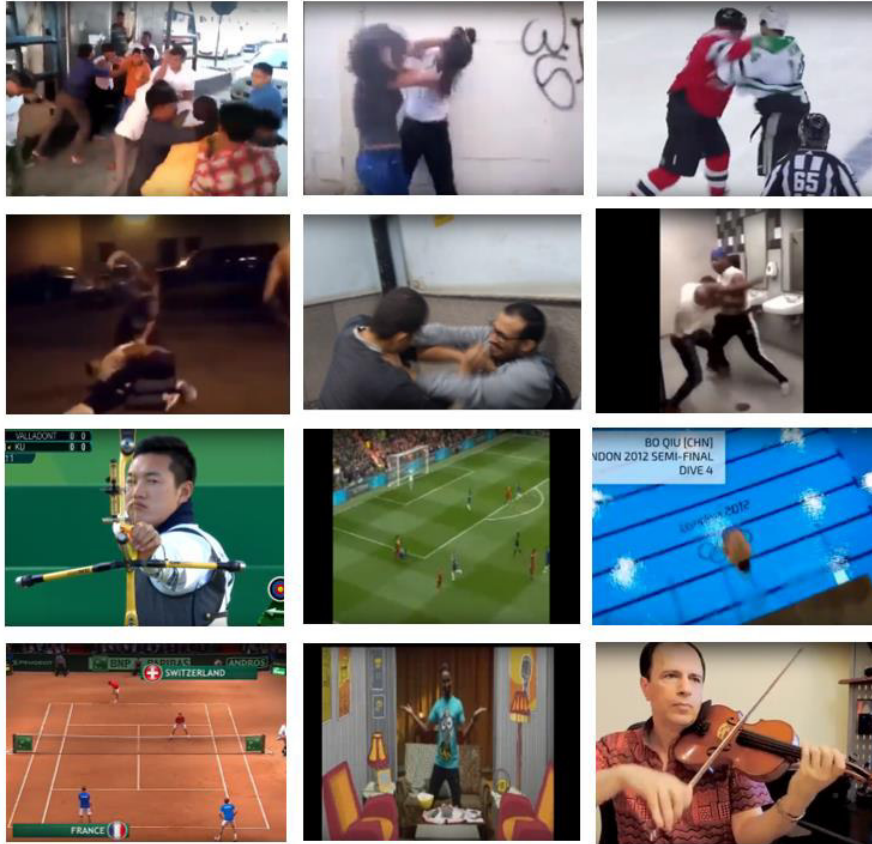


Figura 13 Real Life Violence Dataset

4.2. Componentes

4.2.1. Red neuronal convolucional Xception

El componente de la red neuronal convolucional Xception se usa para extraer características significativas de las secuencias de imágenes obtenidas de los videos de sistemas de vigilancia. Este modelo de red neuronal de convolución ha demostrado un rendimiento de vanguardia debido a un uso más eficiente de sus parámetros.

Xception está basada completamente en capas de convolución separables en profundidad, su arquitectura tiene 36 capas convolucionales que forman la base de extracción de características de la red, las 36 capas convolucionales están estructuradas en 14 módulos, todos los cuales tienen conexiones residuales lineales a su alrededor, excepto el primer y el último módulo.

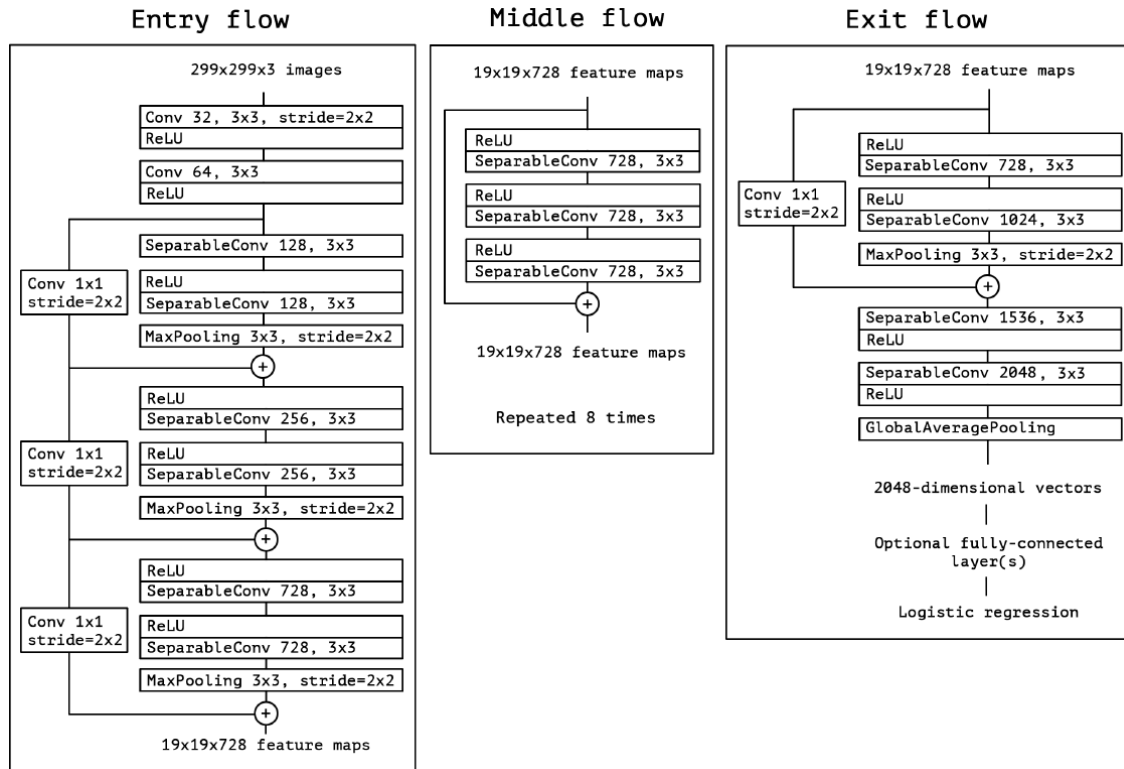


Figura 14 Arquitectura de CNN Xception

Podemos observar la forma de las imágenes esperadas como entrada por el modelo Xception pre-entrenado serán imágenes de la forma $299 \times 299 \times 3$

4.2.2. Long short-term memory (LSTM)

EL componente LSTM tiene un tamaño de 512 neuronas utilizadas para extracción de características temporales, las características se extraen a lo largo de la secuencia de imágenes obtenidas de los videos de sistemas de vigilancia. El siguiente grafico ilustra los detalles de la arquitectura de LSTM.

La forma de entrada del LSTM es (None, 20, 2048), 20 representa el número de secuencias de imágenes que se extraerá de los videos del conjunto de datos de Situaciones de Violencia en la Vida Real tiene y 2048 es el tamaño del vector con los valores de transferencia obtenidos de la red convolucional Xception.

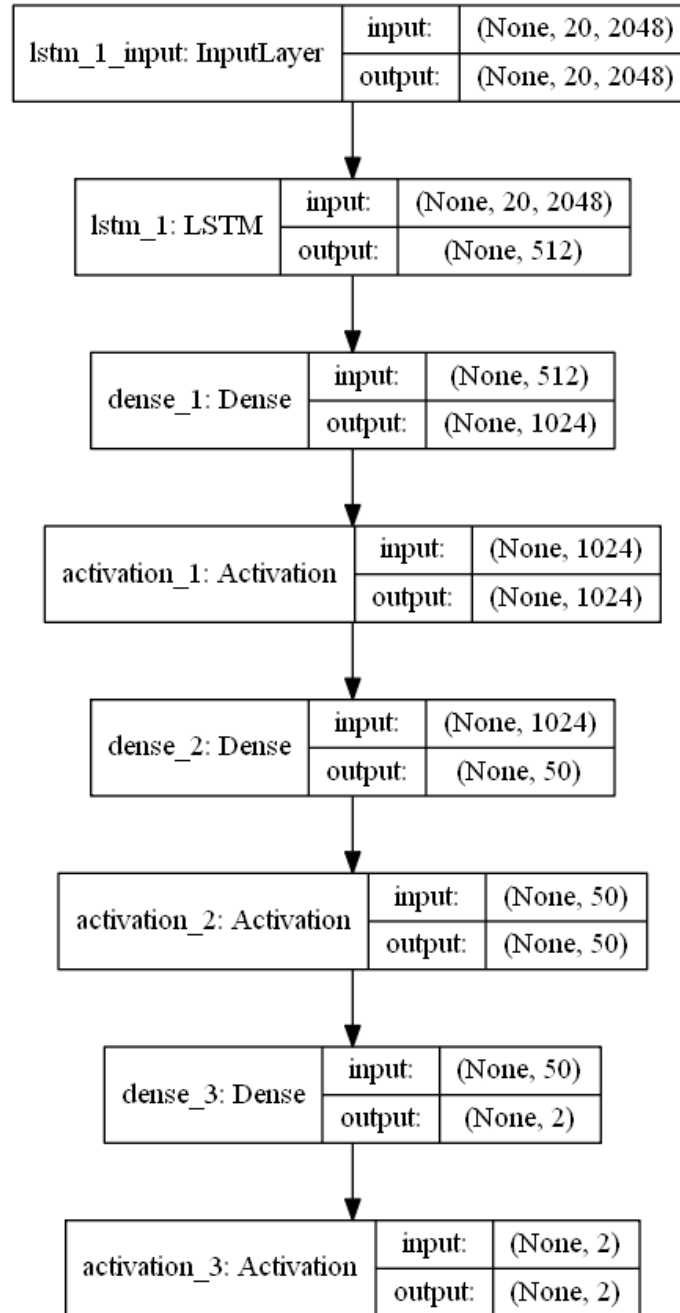


Figura 15 Arquitectura LSTM

4.2.3. Capas totalmente conectadas

En este componente hay 3 capas totalmente conectadas que son utilizadas para la clasificación de la existencia o no de violencia física interpersonal en las secuencias de imágenes de videos del conjunto de datos de Situaciones de Violencia en la Vida Real.

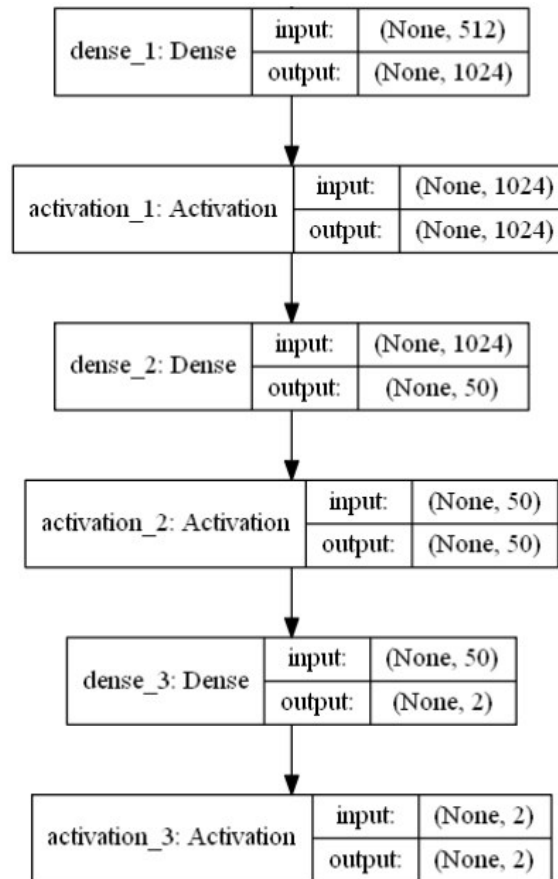


Figura 16 *Arquitectura de Capas Totalmente Conectadas*

El número de neuronas en la primera capa totalmente conectada es 1024 con su función de activación ReLU, en la segunda capa es 1024 con su activación Sigmoidea y en la última capa es 2 porque el número de clases es 2 (Violencia y no violencia) con su activación SoftMax.

4.3. Flujo de trabajo de entrenamiento del modelo de red neuronal propuesto para la detectar violencia física interpersonal en videos.

El modelo propuesto en esta investigación para la detección de violencia física interpersonal en videos realizara un flujo de trabajo para su entrenamiento, como la preparación del Dataset de videos, preprocesamiento de los videos y el entrenamiento del modelo de red neuronal. En la preparación del Dataset, se obtiene un conjunto de videos listo para su preprocesamiento. Mientras en el preprocesamiento de los videos, se extrae los valores de transferencia de las secuencias de

imágenes de los videos mediante el modelo de red convolucional Xception. Finalmente, para el entrenamiento del modelo de red neuronal, los valores de transferencia obtenidos previamente pasan por la red Neuronal Recurrente LSTM, para la extracción de características temporales seguida de dos capas ocultas y la capa de salida es una capa de 2 neuronas con activación softmax, lo que nos da la clasificación final.

A continuación, se muestra el flujo de trabajo a alto nivel de la presente investigación:

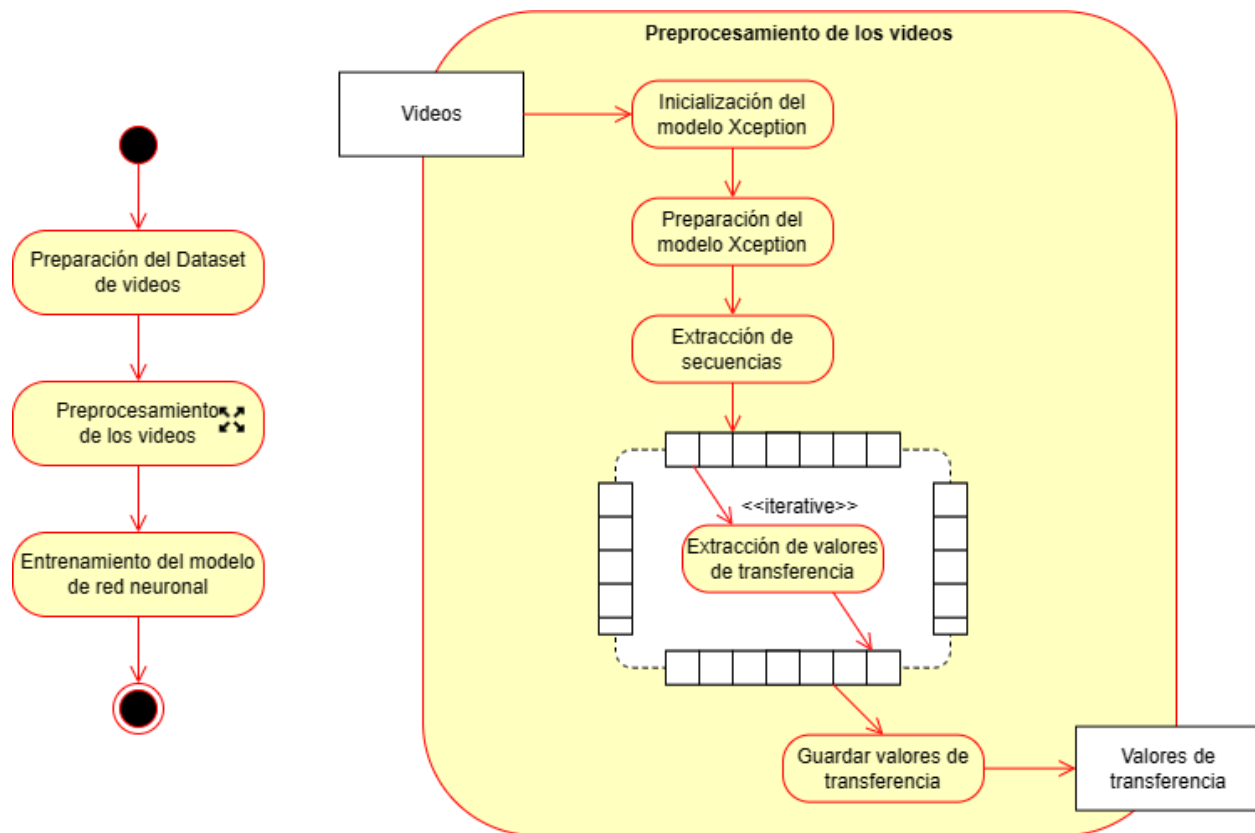


Figura 17 Diagrama de flujo de trabajo de entrenamiento del modelo de red neuronal

4.3.1. Preparación del Dataset de videos

El conjunto de datos obtenido de KAGGLE, tienes dos subconjuntos, teniendo un subconjunto de videos sin violencia y otro conjunto de videos con violencia, este último subconjunto tuvo un

proceso manual de limpieza, validación y recorte, revisando uno a uno los videos. Los pasos seguidos fueron los siguientes:

- Visualizar el video.
- Validar la existencia de violencia.
- Recortar desde el instante donde existe violencia.
- Sobrescribir el video con el mismo nombre.

En el subconjunto de datos de videos con violencia se encontraron algunos videos sin violencia o duplicados, siendo reemplazados por otro recorte no utilizado. Se termino este proceso manual con el nuevo subconjunto de datos guardados en Amazon S3.

4.3.2. Preprocesamiento de los videos

Los videos del nuevo conjunto de datos vienen etiquetados para poder ser diferenciados, los nombres de los videos que empiezan con “V_” son aquellos donde existe violencia y en cambio si el nombre del video empieza con “NV” no existe violencia. Una vez identificados los videos con sus respectivas etiquetas pasamos a su preprocesamiento.

- Dividir el video en secuencia de imágenes: del conjunto de datos dividimos cada video en 20 secuencia de imágenes.
- Redimensionar el tamaño de las secuencias de imágenes extraídos: los cuadros se redimensionan a $(299 * 299 * 3)$ porque la red neuronal Xception requiere imágenes de ese tamaño.
- Shuffle Data: los datos se barajan para evitar que el modelo aprenda un cierto patrón de datos.

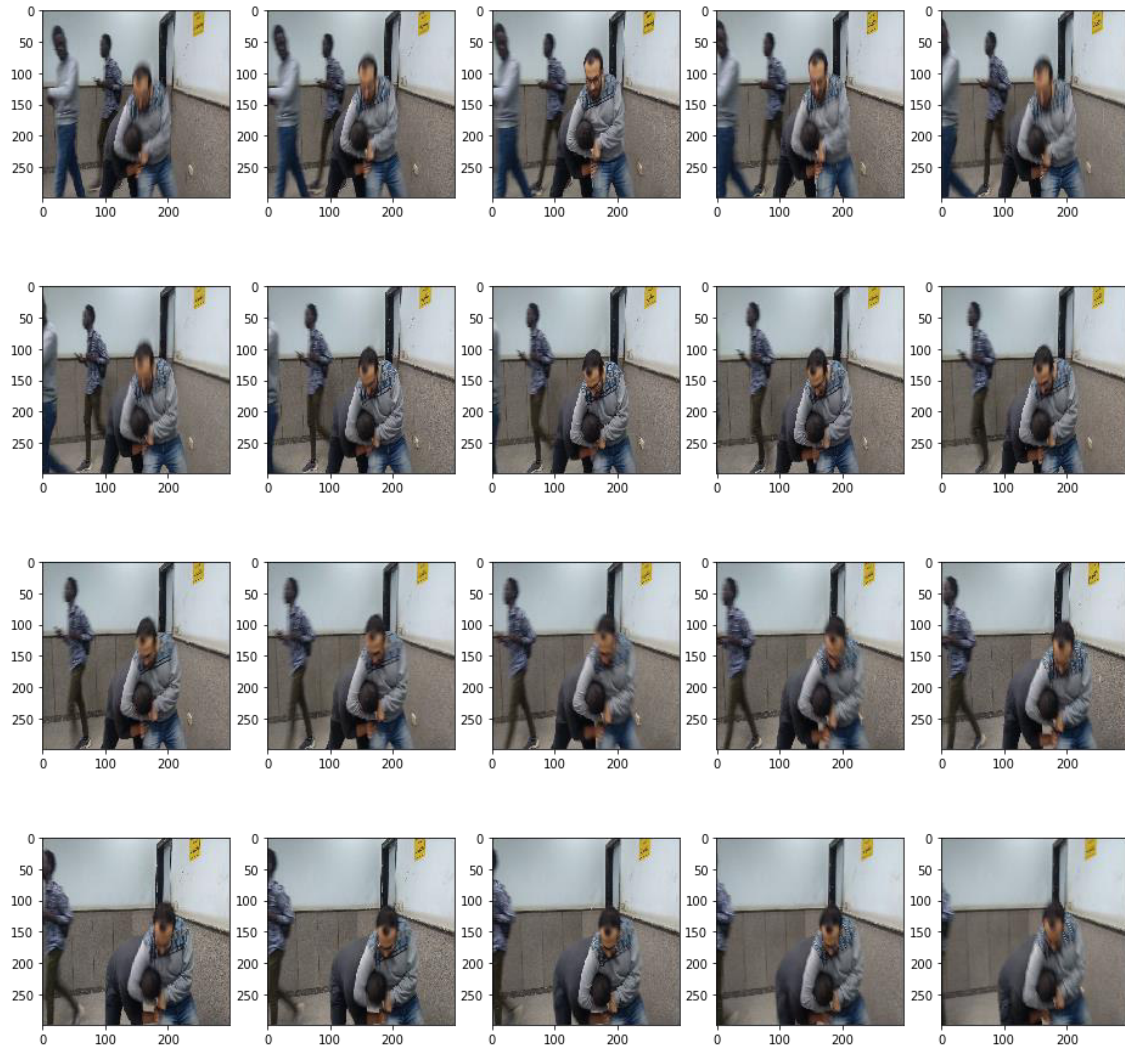


Figura 18 Secuencias de imágenes preprocesadas y extraídas del video

- **Extracción de características**

Se ingresa y procesa en lote 20 secuencias de imágenes extraídas del video, pasando por las diferentes capas del modelo de red convolucional Xception desde la primera capa “input_1” hasta llegar a la capa totalmente conectada “avg_pool”, obteniéndose las características significativas de la capa intermedia “predictions” de cada secuencia de imagen.

Estas características significativas vendrían a ser los valores de transferencia que son usados como la entrada a la red Neuronal LSTM. De cada secuencia de imagen, la red Xception obtiene como

salida un vector de 2048 valores de transferencia. De cada video estamos procesando 20 secuencias de imágenes, por lo que tendremos 20 x 2048 valores de transferencia por video.

- Primera capa(input_1)

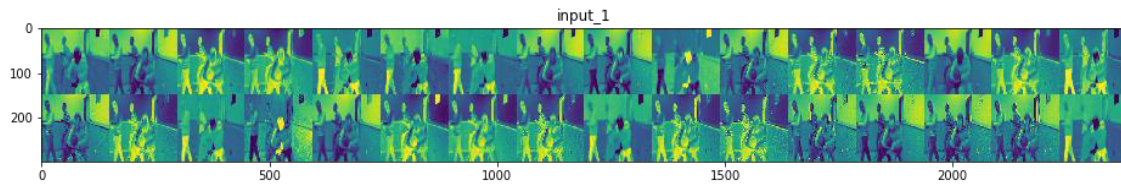


Figura 19 Características extraídas de una imagen en la primera capa de CNN Xception

- Capa de pooling (avg_pool)

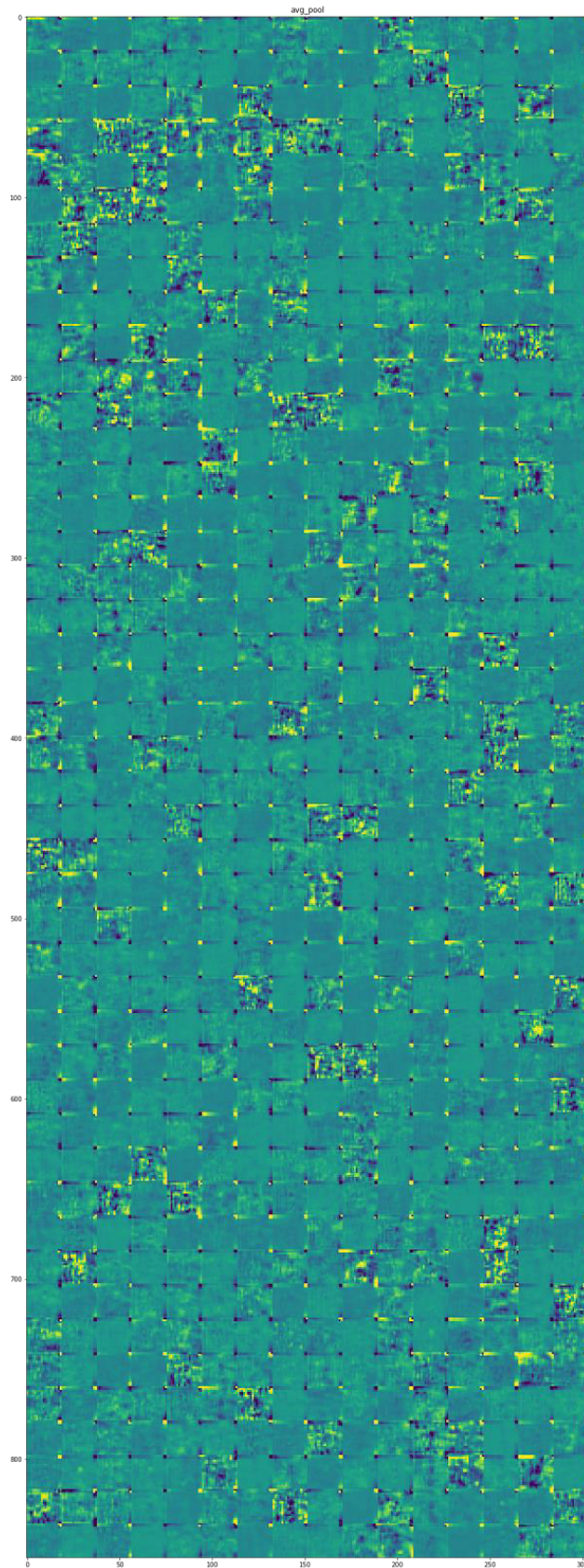


***Figura 20** Características extraídas de una imagen en la capa intermedia avg_pool de CNN Xception*

4.3.3. Entrenamiento del modelo de red neuronal

Para la clasificación entrenaremos la segunda red neuronal usando las clases del conjunto de datos de violencia (Violencia, No Violencia), para que la red aprenda a clasificar imágenes en función de los valores de transferencia del modelo Xception.

La clasificación debe hacerse teniendo en cuenta las 20 secuencia de imágenes obtenidas del video. Si alguno de ellos detecta violencia, el video se clasificará como violento.

4.4. Metodología de desarrollo e implementación del prototipo que detecta violencia física interpersonal

Se hará uso del enfoque del modelo C4 para guiar el diseño e implementación del prototipo que detecta violencia física interpersonal. El modelo C4 proporciona una estructura conceptual que facilita la comprensión y comunicación de la arquitectura del software. Utilizando este enfoque, se podrán definir de manera clara y concisa los distintos componentes del prototipo, así como las relaciones entre ellos, incluyendo los aspectos más relevantes.

4.4.1. Descripción general del prototipo

El prototipo es un servicio web de detección de violencia física interpersonal en secuencias de imágenes de video a través del protocolo de comunicación REST. Aunque puede parecer un prototipo sencillo, en realidad tiene diversos componentes que realizar diversas tareas e integraciones, desde la una API REST, manejo de programación paralela, integraciones con SDK de Amazon Web Services y manejo de estructura de datos.

4.4.2. Arquitectura de contexto del prototipo

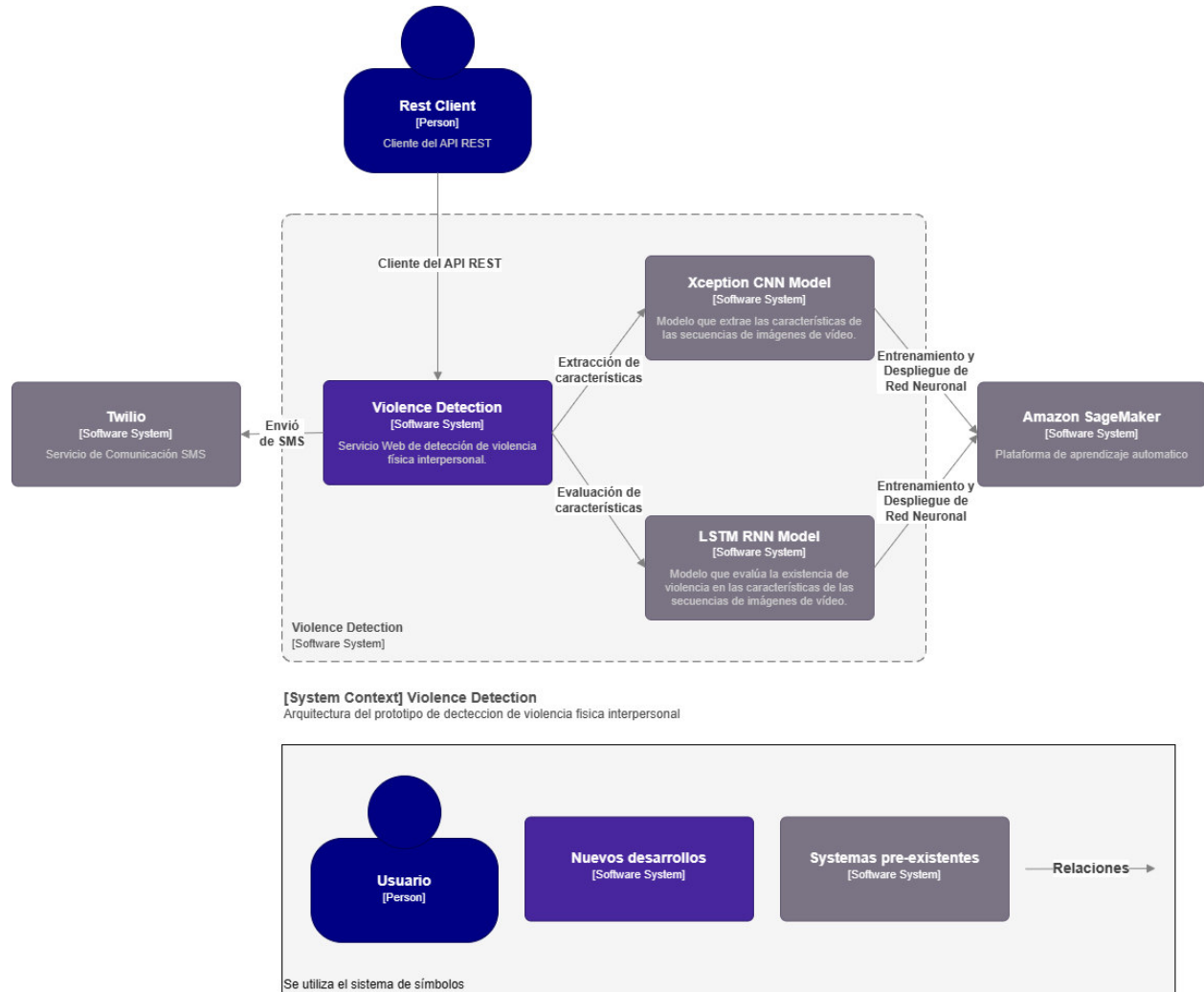


Figura 21 Diagrama de contexto del prototipo

Violence Detection es el prototipo API REST para la detección de violencia física interpersonal, el cual se puede acceder utilizando las solicitudes HTTP. Este prototipo tiene una serie de componentes que se integran con otros sistemas tanto internos como externos, los cuales son:

- Xception CNN Model: Servicio de infraestructura en la nube, desde el cual se extrae las características de las secuencias de imágenes de vídeo.

- LSTM RNN Model: Servicio de infraestructura en la nube, desde el cual se evalúa la existencia de violencia en las características de las secuencias de imágenes de vídeo.
- Amazon SageMaker: Servicio de Infraestructura Cloud de AWS para el entrenamiento y despliegue de las redes neuronales.
- Twilio: Representa el API de Twilio para el envío de notificaciones SMS.

4.4.3. Arquitectura Lógica

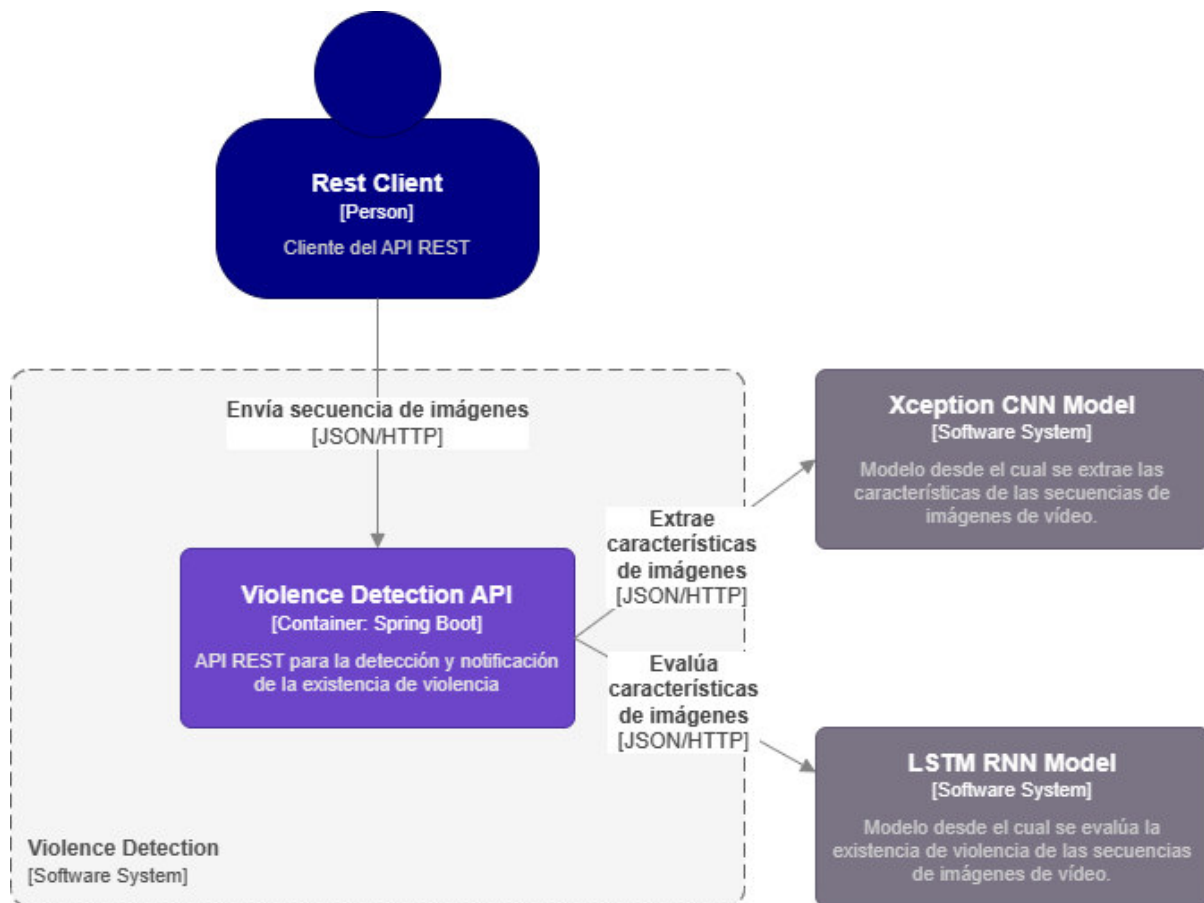
- **Arquitectura de contenedores**

Violence Detection es un prototipo para la detección de violencia física interpersonal, compuesto por un principal componente que juega un papel importante en la arquitectura. Este componente es el backend que tiene una importante relevancia para poder procesar las secuencias de imágenes, orquestar los servicios de los modelos de redes neuronales, envío de notificaciones SMS, etc. En ese sentido Violence Detection API es el componente de backend que expone los servicios y lógica de negocio al prototipo.

En la siguiente vista de la arquitectura podemos apreciar los siguientes sistemas:

- Violence Detection API: Corresponde al componente Backend (Spring Boot) que da soporte a los clientes REST mediante la exposición de servicios
- Xception CNN Model: Modelo de red neuronal convolucional, desde el cual se extrae las características de las secuencias de imágenes de vídeo.
- LSTM RNN Model: Modelo de red neuronal recurrente, desde el cual se evalúa la existencia de violencia en las características de las secuencias de imágenes de vídeo.

Violence Detection API es una API REST desarrollado en Spring Boot bajo la arquitectura de microservicios, por lo tanto, cuenta con su propio runtime basado en Apache Tomcat y puede ejecutarse por sí mismo sin necesidad de un servidor de aplicaciones.



[Containers] Violence Detection

Diagrama que muestra los contenedores del prototipo de detección de violencia física interpersonal.

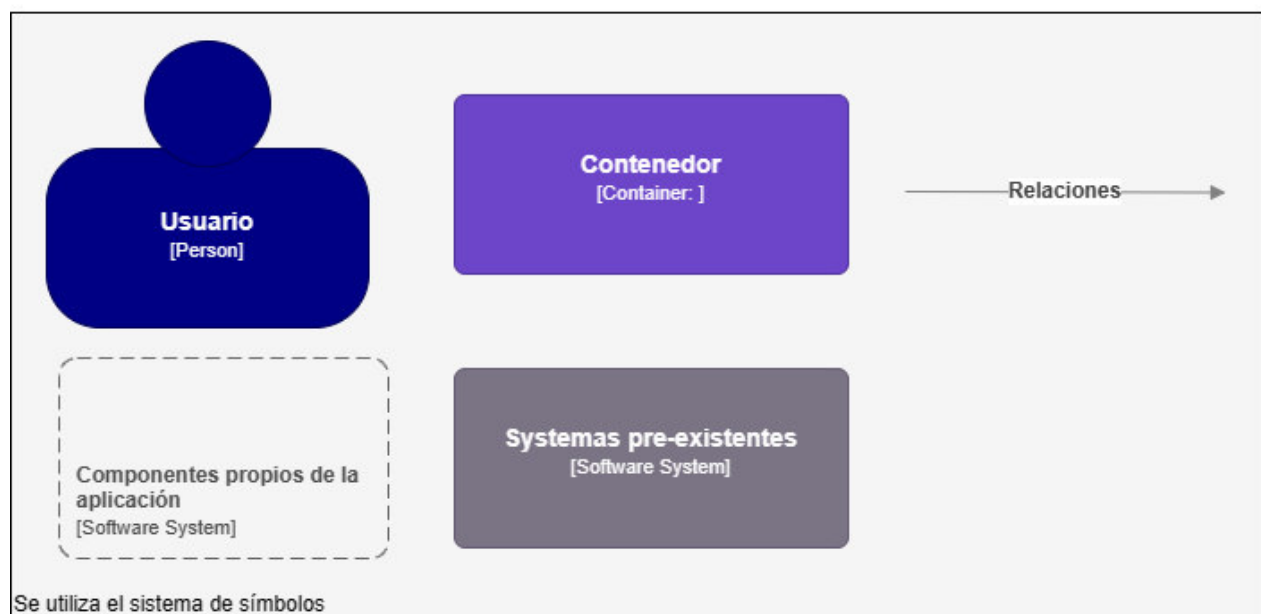


Figura 22 Diagrama de contenedores del prototipo Violence Detection

4.4.4. Arquitectura de componentes

Violence Detection API es construido en capas, donde cada una de ellas cubre una responsabilidad del sistema, estas capas sirven para organizar mejor el proyecto y estandarizar la forma de trabajo.

Las capas definidas son:

- **Controllers:** Representa la capa de servicios y en esta se definen todos los métodos que serán expuestos como servicios REST. Esta capa se encarga de recibir las peticiones y de legarlas a la capa de servicios. Esta capa no procesa nada de lógica de negocios, solo se limita al procesamiento de solicitudes y su serialización de las respuestas en formato JSON.
- **Services:** Capa lógica de la arquitectura, donde se lleva a cabo toda la lógica negocio, lógica de manejo de datos, orquestación de Services e integraciones con SDKs.
- **Clients:** Representa la capa de clientes REST, donde se realiza las integraciones con API REST de terceros.

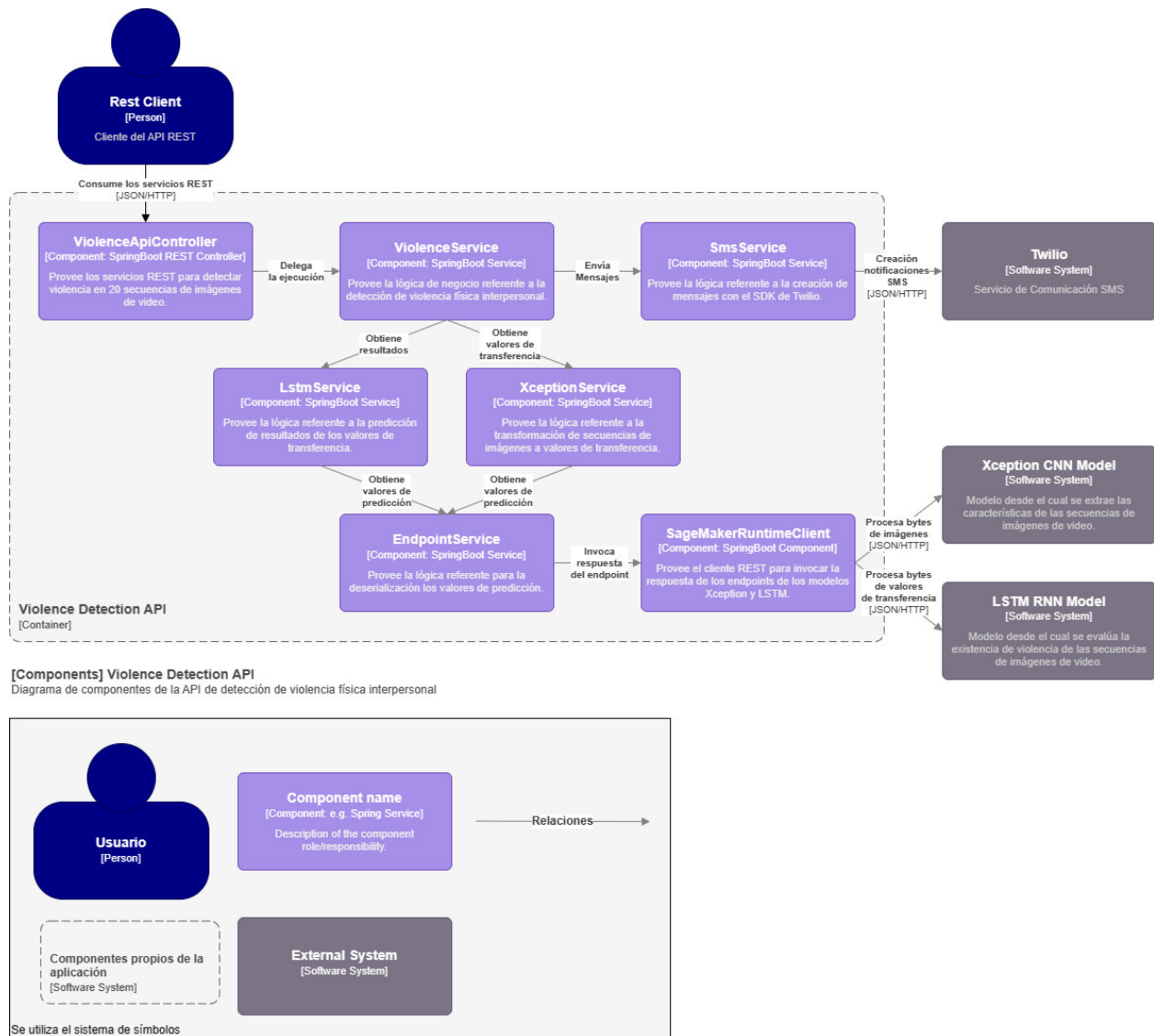


Figura 23 Diagrama de componentes

Para entender mejor el funcionamiento de la API REST, agruparemos en un dominio una serie de clases relacionadas entre sí y que responde en función de una entidad.

- **Dominio de violencia**

El dominio de violencia, esta conformado por las clases ViolenceApiController, ViolenceService, LstmService, XceptionService, EndpointService, SageMakerRuntimeClient, SmsService. Este dominio se centra en la detección de violencia física interpersonal y realizar

notificaciones SMS. En este dominio pueden entrar en juego otras clases como los utilitarios FrameUtils, que se encarga de la transformación de una imagen a un arreglo de N dimensiones y NDArrayUtils que realiza la conversión del arreglo de N dimensional a 4 dimensiones.

4.4.5. Arquitectura de bajo nivel

El siguiente diagrama ilustra perfectamente como está conformado el contenedor Violence Detection API con su dominio de violencia:

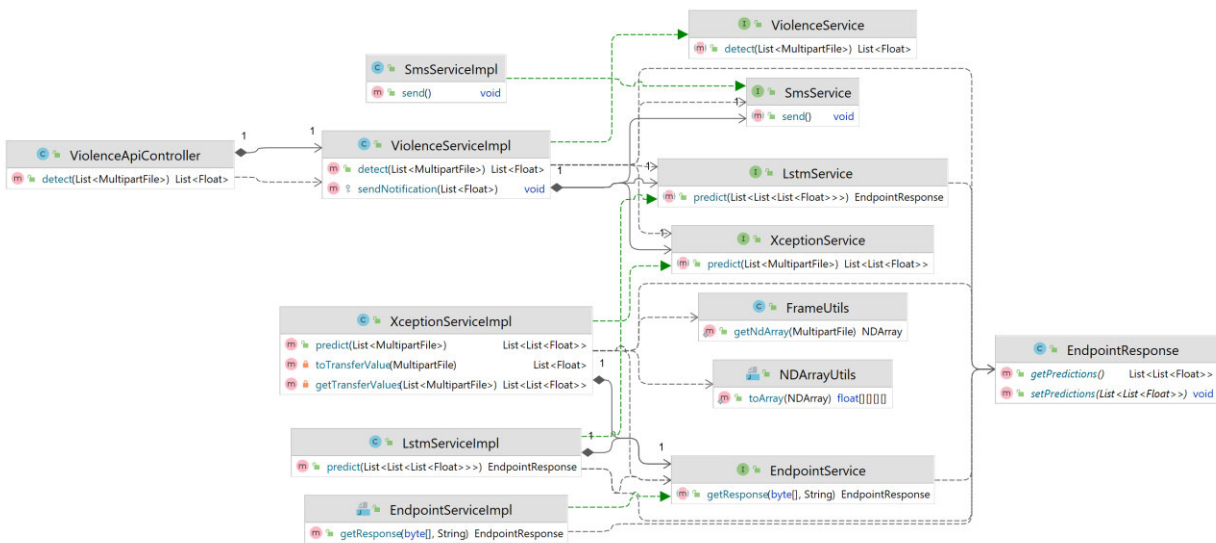


Figura 24 Diagrama de clases del dominio violencia

4.4.6. Arquitectura física

Se ilustra como los contenedores se despliegan físicamente dentro de los servidores de la infraestructura en la nube.

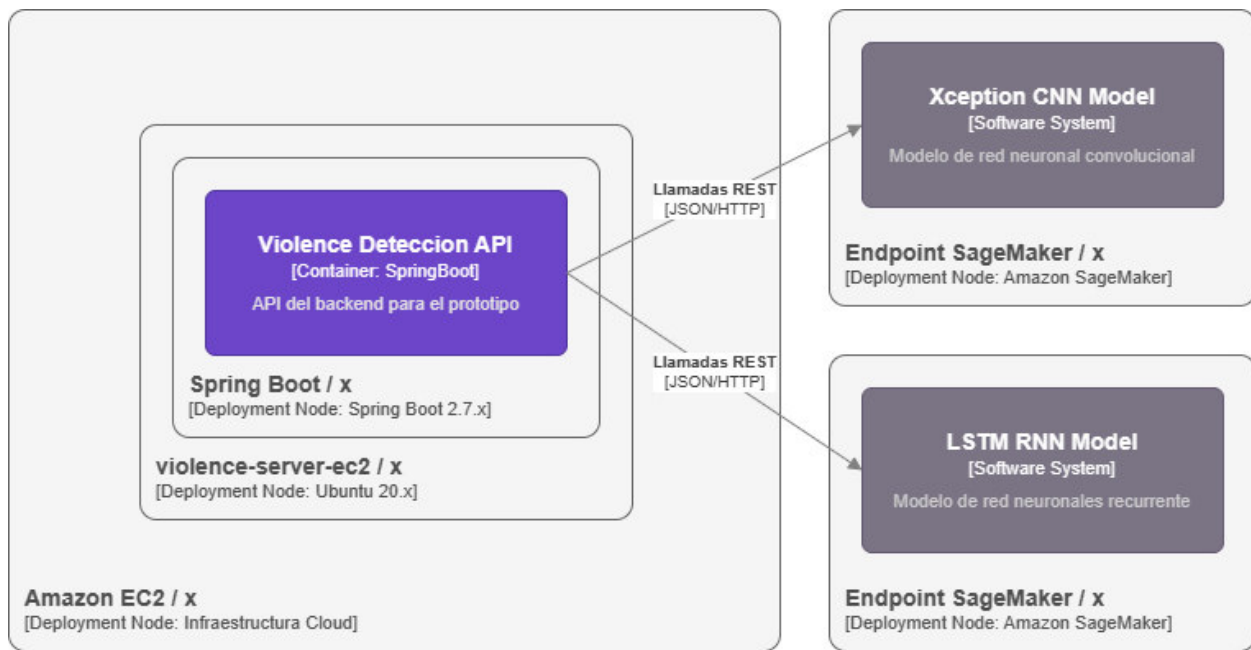


Figura 25 Diagrama de despliegue

El prototipo de Violence Detection cuenta con una robusta infraestructura en la nube que le permite desplegar sus servicios. Para esto el prototipo cuenta con una instancia de Amazon EC2 para desplegar el API REST y Amazon SageMaker para desplegar los modelos de redes neuronales.

Los elementos que aparecerán en este diagrama son:

- **Violence Detection API:** El API REST del prototipo esta desplegado sobre un servidor llamado “violence-server-ec2” alojado en la infraestructura de Amazon EC2, por lo que la gestión y escalamiento de este componente está a cargo de Amazon EC2.
- **Xception CNN Model:** El modelo de red neuronal convolucional esta desplegado sobre Endpoint SageMaker, por lo que la gestión y escalamiento de este componente está a cargo de Amazon SageMaker.

- LSTM RNN Model: El modelo de red neuronal recurrente esta desplegado sobre Endpoint SageMaker, por lo que la gestión y escalamiento de este componente está a cargo de Amazon SageMaker.

Capítulo V: Resultados y discusión

5.1. Resultados

Posterior a extraer las características significativas de las secuencias de imágenes, las guardamos como todos los valores de transferencia de videos en un archivo, para luego poder cargar esos valores de transferencia en la memoria para entrenar la red LSTM.

La configuración del entrenamiento para el modelo es:

- Épocas: 200
- Batch size (Tamaño de lote): 500
- Optimizador: Adam
- Función de activación en la última capa: Softmax
- Función de pérdidas: Entropía cruzada

5.1.1. Dataset

El conjunto de datos de Situaciones de Violencia en la vida real se dividió en datos de entrenamiento que contiene 1600 videos (800 violencia y 800 no violencia), utilizándose 1280 videos para entrenar y 320 videos para validar, también de la división del conjunto de datos se obtuvo los datos de prueba que contiene 400 videos (200 violencia y 200 no violencia). No hay técnicas de aumento de datos, solo se utilizan los videos originales.

Para diferenciar los videos donde existe o no violencia en este conjunto de datos, se pusieron etiquetas al comienzo de los dos primeros caracteres de los nombres de los videos, teniendo a “NV” como aquellos videos sin presencia violencia y “V_” aquellos videos con presencia de violencia. En las siguientes figuras se podrá visualizar una muestra de este conjunto de datos.

Nombre	Tipo	Tamaño	Duración
 NV_1.mp4	Archivo MP4	7,796 KB	00:00:02
 NV_2.mp4	Archivo MP4	518 KB	
 NV_3.mp4	Archivo MP4	893 KB	00:00:05
 NV_4.mp4	Archivo MP4	1,431 KB	00:00:04
 NV_5.mp4	Archivo MP4	1,431 KB	00:00:04
 NV_6.mp4	Archivo MP4	1,323 KB	00:00:04
 NV_7.mp4	Archivo MP4	2,016 KB	00:00:03
 NV_8.mp4	Archivo MP4	1,271 KB	00:00:05
 NV_9.mp4	Archivo MP4	734 KB	00:00:04
 NV_10.mp4	Archivo MP4	1,125 KB	
 NV_11.mp4	Archivo MP4	1,622 KB	00:00:05
 NV_12.mp4	Archivo MP4	799 KB	00:00:05
 NV_13.mp4	Archivo MP4	1,400 KB	00:00:06
 NV_14.mp4	Archivo MP4	2,641 KB	00:00:04
 NV_15.mp4	Archivo MP4	1,326 KB	00:00:05
 NV_16.mp4	Archivo MP4	971 KB	00:00:04
 NV_17.mp4	Archivo MP4	1,591 KB	00:00:05
 NV_18.mp4	Archivo MP4	1,536 KB	00:00:04
 NV_19.mp4	Archivo MP4	1,536 KB	00:00:04
 NV_20.mp4	Archivo MP4	1,669 KB	00:00:05

***Figura 26** Muestra de videos sin presencia de violencia del conjunto de datos*

Nombre	Tipo	Tamaño	Duración
 V_1.mp4	Archivo MP4	3,067 KB	00:00:06
 V_2.mp4	Archivo MP4	3,134 KB	00:00:06
 V_3.mp4	Archivo MP4	1,645 KB	00:00:06
 V_4.mp4	Archivo MP4	1,905 KB	00:00:06
 V_5.mp4	Archivo MP4	1,821 KB	00:00:05
 V_6.mp4	Archivo MP4	1,922 KB	00:00:06
 V_7.mp4	Archivo MP4	11,662 KB	00:00:04
 V_8.mp4	Archivo MP4	14,159 KB	00:00:05
 V_9.mp4	Archivo MP4	4,947 KB	00:00:07
 V_10.mp4	Archivo MP4	1,172 KB	00:00:05
 V_11.mp4	Archivo MP4	1,220 KB	00:00:07
 V_12.mp4	Archivo MP4	705 KB	00:00:04
 V_13.mp4	Archivo MP4	746 KB	00:00:05
 V_14.mp4	Archivo MP4	1,398 KB	00:00:06
 V_15.mp4	Archivo MP4	754 KB	00:00:03
 V_16.mp4	Archivo MP4	598 KB	00:00:03
 V_17.mp4	Archivo MP4	2,649 KB	00:00:06
 V_18.mp4	Archivo MP4	1,848 KB	00:00:04
 V_19.mp4	Archivo MP4	993 KB	00:00:03
 V_20.mp4	Archivo MP4	1,807 KB	00:00:04

***Figura 27** Muestra de videos con presencia de violencia del conjunto de datos*

5.1.2. Trazo del entrenamiento y métricas

Para el entrenamiento y validación del modelo se utilizaron 1280 y 320 muestras respectivamente, en la siguiente tabla se mostrará las métricas obtenidas en el proceso por épocas.

- **loss**, es el valor de pérdida del modelo con los datos de entrenamiento en determinada época del proceso de entrenamiento
- **acc**, es el valor de precisión del modelo con los datos de entrenamiento en determinada época del proceso de entrenamiento
- **val_loss**, es el valor de pérdida del modelo con los datos de validación en determinada época del proceso de entrenamiento
- **val_acc**, es el valor de precisión del modelo con los datos de validación en determinada época del proceso de entrenamiento

***Tabla 4** Trazo de entrenamiento*

epoca	loss	acc	val_loss	val_acc
1 / 200	0.339175706	0.49866666	0.27708259	0.50705881
2 / 200	0.250140737	0.61466666	0.22152647	0.57529413
3 / 200	0.188351174	0.67599998	0.1520205	0.7917647
4 / 200	0.139320513	0.80533336	0.10250642	0.8529412
5 / 200	0.096939221	0.87600001	0.0846689	0.88000001
6 / 200	0.072594342	0.90133333	0.07914569	0.9035294
7 / 200	0.069402717	0.90399998	0.06056787	0.91058822
8 / 200	0.060734645	0.92	0.05173317	0.92941176
9 / 200	0.04213484	0.94133333	0.06743423	0.91176472
10 / 200	0.041682915	0.94799999	0.04563038	0.93764707
11 / 200	0.035759607	0.95333332	0.04612509	0.93647056
12 / 200	0.031189591	0.95866666	0.0461804	0.94588234
13 / 200	0.024290395	0.972	0.04159006	0.94235294
14 / 200	0.018264037	0.98	0.03546452	0.95647056
15 / 200	0.017110857	0.97600001	0.03432972	0.95529413

16 / 200	0.01121127	0.98666666	0.03958429	0.94470587
17 / 200	0.009878978	0.988	0.03400405	0.95882353
18 / 200	0.006991586	0.99199998	0.03703899	0.95176471
19 / 200	0.005803175	0.99466666	0.03173187	0.96117645
20 / 200	0.003238211	0.99733335	0.03642868	0.95411766
21 / 200	0.002295593	0.99866668	0.0363005	0.95176469
22 / 200	0.000915127	1	0.03668088	0.95411764
23 / 200	0.000388329	1	0.03761878	0.95411764
24 / 200	0.000215813	1	0.03909282	0.95176471
25 / 200	0.000156854	1	0.03956679	0.95176471
26 / 200	0.000117616	1	0.03972483	0.95411764
27 / 200	8.72E-05	1	0.04086795	0.95058822
28 / 200	7.46E-05	1	0.04221132	0.94941175
29 / 200	6.77E-05	1	0.04305487	0.94941176
30 / 200	6.26E-05	1	0.04340715	0.94823531
31 / 200	5.77E-05	1	0.0434535	0.94823531
32 / 200	5.25E-05	1	0.04340842	0.94823529
33 / 200	4.90E-05	1	0.04332096	0.95058822
34 / 200	4.58E-05	1	0.04325588	0.95058822
35 / 200	4.39E-05	1	0.04320595	0.95058822
36 / 200	4.22E-05	1	0.04318814	0.95058822
37 / 200	4.08E-05	1	0.04319121	0.95058822
38 / 200	3.95E-05	1	0.04323606	0.95058822
39 / 200	3.82E-05	1	0.04327503	0.95058822
40 / 200	3.73E-05	1	0.04331587	0.95058822
...	...	...	...	...
41 / 200	3.65E-05	1	0.04334659	0.95058822
198 / 200	1.86E-05	1	0.04435877	0.95411764
199 / 200	1.86E-05	1	0.04436704	0.95411764
200 / 200	1.85E-05	1	0.04437534	0.95411764

Los siguientes dos gráficos representan la exactitud y pérdida en el proceso entrenamiento del modelo por épocas, respectivamente.

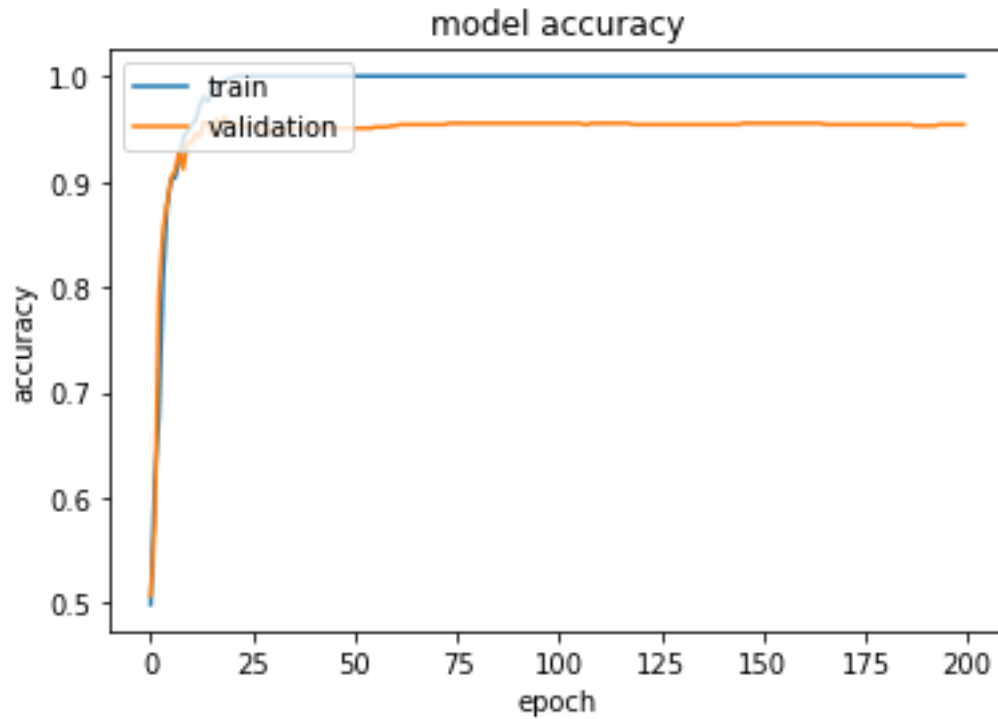


Figura 28 Gráfico de precisión del Modelo

En la Figura 28 se observa que se logra obtener un valor precisión de 100% con los datos de entrenamiento sin mucho sobreajuste en el modelo, de igual manera con los datos validación, pero con un valor de precisión más bajo de 95.41%. También se logra ver que a partir de la época 27 se obtiene resultados precisión aceptables y estables a medida que va aumentando las épocas en los datos de entrenamiento y validación.

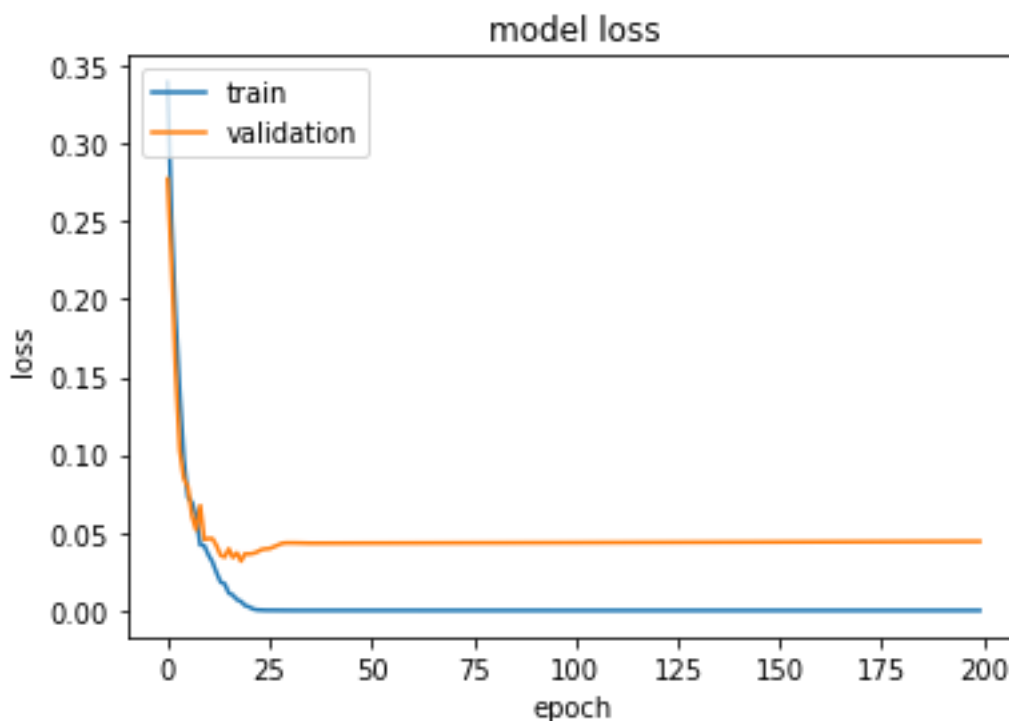


Figura 29 Gráfico de pérdidas del Modelo

En la Figura 29 se observa que se logra obtener un valor de pérdida de 0.001% con los datos de entrenamiento sin mucho sobreajuste en el modelo, de igual manera con los datos validación, pero con un valor de pérdida más alto de 4.43%. También se logra ver que a partir de la época 38 se obtiene resultados del valor de pérdida más aceptables y estables a medida que va aumentando las épocas en los datos de entrenamiento y validación.

5.1.3. Prueba del modelo y métricas

En la prueba del modelo se utilizó el 20% del total de videos del conjunto de datos que serían 400 videos (200 violencia y 200 no violencia). Estos videos no se han utilizado para entrenar el modelo de la red neuronal.

Se obtuvo los siguientes resultados grosso modo:

- En la evaluación del modelo el valor calculado de la función de pérdida que se logró obtener fue de **5.98%**
- En la evaluación del modelo el valor calculado de la exactitud del modelo fue de **93.5%**

Para describir de mejor el desempeño de nuestro modelo de visión artificial para la detección de violencia física interpersonal, usaremos una matriz confusión en el conjunto de datos prueba mencionado en esta sección que se muestra en la Figura 30.

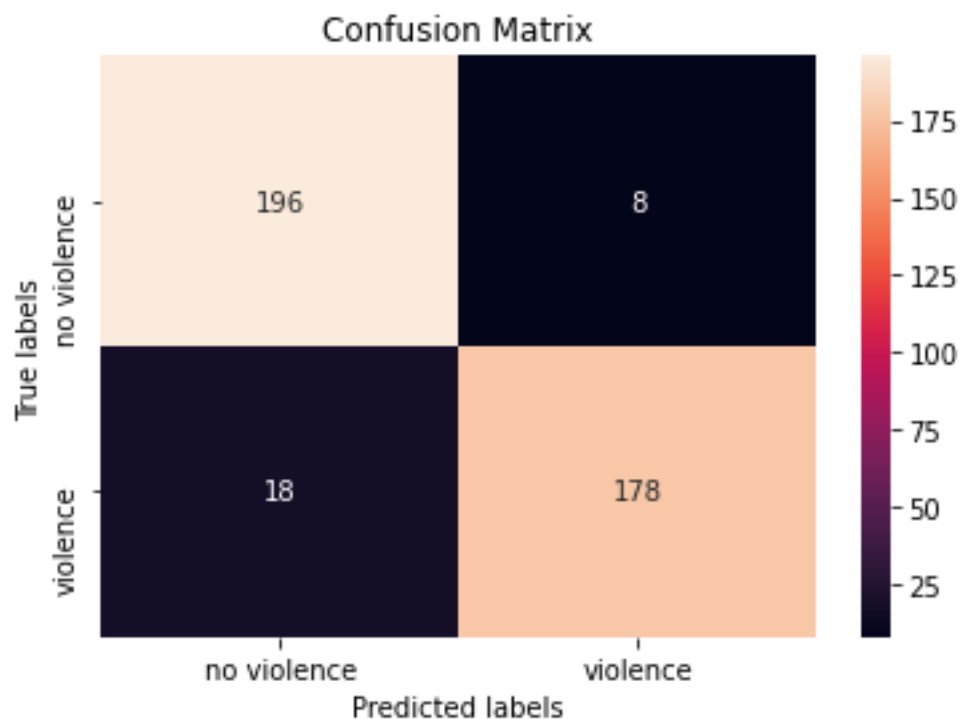


Figura 30 Matriz de confusión

Ya con la matriz de confusión podremos hallar e interpretar mejor las siguientes métricas:

- Sensibilidad o recall: nos da una probabilidad de **90.81%** de que cuando exista violencia en un video, el modelo la clasifique así.

- Precisión: nos da una probabilidad de **95.69%** de predecir la existencia violencia en un video cuando realmente lo sea.
- Exactitud: nos da una probabilidad de **93.5%** de aciertos de nuestro modelo en la clasificación de la existencia o no de violencia en un video.

5.2. Discusión

La selección de la red neuronal convolucional Xception se realizó mediante la evaluación del rendimiento en el conjunto de datos Hockey Fight dataset, la Tabla 6.2 proporciona los valores de exactitud obtenidos para los diversos conjuntos de datos encontrados en las investigaciones (Sudhakaran & Lanz, 2017), (T. Zhang et al., 2018) y (Ben Mabrouk & Zagrouba, 2017) con sus técnicas de vanguardia, estas investigaciones están relacionadas de manera directa con la detección de violencia en videos. De la tabla, se puede ver que el método propuesto obtiene una exactitud de 98% mejorando los resultados de las técnicas propuestas en (Sudhakaran & Lanz, 2017), (T. Zhang et al., 2018) y (Ben Mabrouk & Zagrouba, 2017) con una exactitud de $97.1 \pm 0.55\%$, $96.5 \pm 1.04\%$ y $88.6 \pm 1.2\%$ respectivamente en el caso del conjunto de datos Hockey Fight dataset.

Tabla 5 Comparación de resultados de clasificación

Dataset	Metodo	N° de Videos		Exactitud
Violent-Flows Dataset / Crowd Violence Dataset	ConvLSTM [11]	246	123 violencia	$94.57 \pm 2.34\%$
	Semi-supervised learning framework [69]			$92.25 \pm 0.12\%$
	DiMOLIF [70]		123 no violencia	85.83 ± 4.26
	Xception + LSTM	—		
Movies Dataset	ConvLSTM [11]	200	100 violencia	$100 \pm 0\%$

			100 no violencia	
	Semi-supervised learning framework [69]	—		
	DiMOLIF [70]	—		
	Xception + LSTM	—		
BEHAVE Dataset	ConvLSTM [11]	—		
	Semi-supervised learning framework [69]	80	20 violencia	88.26 ± 0.11%
			60 no violencia	
	DiMOLIF [70]	—		
	Xception + LSTM	—		
Hockey Fight Dataset	ConvLSTM [11]	500	violencia	97.1 ± 0.55%
	Semi-supervised learning framework [69]		no violencia	96.5 ± 1.04%
	DiMOLIF [70]			88.6 ± 1.2%
	Xception + LSTM	1000	500 violencia	98%
			500 no violencia	
Real Life Violence Situations Dataset	ConvLSTM [11]	—		
	Semi-supervised learning framework [69]	—		
	DiMOLIF [70]	—		
	Xception + LSTM	2000	1000 violencia	93.5%
			1000 no violencia	

En la Tabla 6.2 tenemos a (T. Zhang et al., 2018) quienes desarrollaron un algoritmo de detección de violencia efectivo mediante un algoritmo de aprendizaje de diccionario semi-supervisado que integra el aprendizaje de diccionarios y la capacitación de clasificadores, donde observamos que para el caso del conjunto de datos BEHAVE Dataset utiliza 80 videos de los cuales 20 son videos de violencia y 60 son de no violencia, por lo cual los datos de entrenamiento no están balanceados y si tenemos muy pocos videos de violencia, afectará en el aprendizaje del algoritmo para detectar un video con violencia, en cambio en nuestro conjunto de datos elegido Real Life Violence

Situations dataset con 2000 videos de los cuales 1000 son videos de violencia y 1000 son de no violencia, los datos de entrenamiento están balanceados.

Si de la Tabla 6.2 solo comparamos en valor de exactitud obtenido por cada investigación tendríamos que nuestra propuesta supera todos los resultados de (Ben Mabrouk & Zagrouba, 2017), quienes proponen un método para la detección de violencia, donde teniendo un vídeo de entrada se detecta primero un conjunto de puntos STIP para determinar los marcos de interés que se utilizan para extraer un descriptor, luego, finalmente las características extraídas a través de las función DiMOLIF basada en puntos STIP e información de flujo óptico se utilizan como una entrada para el clasificador SVM, ya que nosotros obtenemos valores de exactitud de 98% y 93.5% en los conjuntos de datos Violent-Flows dataset y Real Life Violence Situations Dataset respectivamente a diferencia de (Ben Mabrouk & Zagrouba, 2017) que solo obtienen 85.83 ± 4.26 y $88.6 \pm 1.2\%$. en los conjuntos de datos Violent-Flows dataset y Hockey Fight dataset respectivamente.

Finalmente, de la Tabla 6.2 siguiendo la comparación del valor de exactitud obtenido por cada investigación tendríamos que (Sudhakaran & Lanz, 2017), quienes desarrollaron un modelo de red neuronal profunda entrenable de extremo a extremo para detectar videos violentos, que consta de capas convolucionales, normalización y agrupación (azul) y el Hidden State, obtuvieron mejores resultados en los valores de exactitud obviando el conjunto de datos de Hockey Fight dataset de $100 \pm 0\%$ y 96.4% en los conjuntos de datos Movies dataset y Violent-Flows dataset respectivamente, en contraste con la nuestra que solo obtenemos un 93.5% en el conjunto de datos Real Life Violence Situations Dataset. Pero a diferencia de (Sudhakaran & Lanz, 2017) y en (T. Zhang et al., 2018) (Ben Mabrouk & Zagrouba, 2017) que utilizaron los conjuntos de datos Violent-Flows Dataset, Movies Dataset, BEHAVE Dataset, Hockey Fight Dataset con 246, 200,

80, 500 videos respectivamente, nosotros utilizamos una mayor cantidad de videos en nuestros conjuntos de datos de Violent-Flows dataset y Real Life Violence Situations Dataset con 1000 y 2000 videos respectivamente, otra diferencia importante en esta investigación es que el conjunto de datos Real Life Violence Situations Dataset contiene gran variedad de situaciones de violencia en la vida real, que simula mejor lo captado por las cámaras en los sistemas de videovigilancia.

En (Roa et al., 2018) también está relacionada de manera directa con la detección de violencia, pero a diferencia de nuestro trabajo que realizamos la detección de violencia en videos ellos proponen un componente de software prototipo para la detección automática de situaciones de inseguridad basadas en el reconocimiento de voz, el procesamiento del lenguaje natural y el reconocimiento de patrones, obtuvieron buenos resultados en ambientes controlados, pero en ambientes reales dejaron mucho que desear debido a la calidad de entrada de audio, porque los micrófonos de las cámaras de seguridad no son de buena calidad y normalmente se encuentran en lugares que no son muy accesibles, lo que dificulta el reconocimiento de voz, este problema no se presenta en nuestro trabajo porque nosotros analizamos las secuencias de imágenes de los videos obviando la calidad de audio que no es un factor importante para nosotros.

5.3. Conclusiones

- Se diseño e implementó una arquitectura con un modelo de red neuronal convolucional Xception y LSTM para la detección de violencia física interpersonal que se podría en sistemas de videovigilancia.
- Se utilizo un conjunto de datos de video con situaciones de violencia en la vida real, para entrenar el modelo de red neuronal convolucional Xception con LSTM, que a diferencia de las investigaciones citadas en nuestro trabajo sus conjuntos de datos no tienen gran

variedad de videos que simulen lo captado por una cámara de un sistema de vigilancia en la vida real.

- Se desarrollo un módulo de entrenamiento para el modelo de red neuronal convolucional Xception con LSTM, para que a medida que aumente el conjunto de videos analizados en el módulo, también aumente el conjunto de datos de entrenamiento etiquetados correctamente.
- Validamos la arquitectura del modelo de red neuronal convolucional Xception con LSTM para la detección de violencia física interpersonal, donde se obtuvo buenos resultados con un gran conjunto de videos utilizados a diferencia de las investigaciones citadas en este trabajo.

5.4. Trabajos futuros

Para trabajos futuros, se buscará que nuestro modelo de visión artificial que además de detectar la violencia física interpersonal en videos, ver la intensidad de la violencia física interpersonal ejercida, también clasificar el tipo de violencia que existe en el video.

Referencias

- Adam, A., Rivlin, E., Shimshoni, I., & Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3), 555–560. <https://doi.org/10.1109/TPAMI.2007.70825>
- Ahmed, F., Topu, B. A., & Islam, S. M. M. (2019). HOG and Gabor Filter Based Pedestrian Detection using Convolutional Neural Networks. *2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019*, 1–6. <https://doi.org/10.1109/ECACE.2019.8679133>
- Ahonen, T., Hadid, A., & Pietik, M. (2004). Face Recognition with Local Binary Patterns. *European Conference on Computer Vision*, 469–481. https://doi.org/https://doi.org/10.1007/978-3-540-24670-1_36
- Babiker, M., Khalifa, O. O., Htike, K. K., Hassan, A., & Zaharadeen, M. (2018). Automated daily human activity recognition for video surveillance using neural network. *2017 IEEE International Conference on Smart Instrumentation, Measurement and Applications, ICSIMA 2017*, 1–5. <https://doi.org/10.1109/ICSIMA.2017.8312024>
- Bautista-Durán, M., García-Gómez, J., Gil-Pita, R., Mohíno-Herranz, I., & Rosa-Zurera, M. (2017). Energy-efficient acoustic violence detector for smart cities. *International Journal of Computational Intelligence Systems*, 10(1), 1298–1305. <https://doi.org/10.2991/IJCIS.10.1.89>
- Beliz, G. (2012). Gobernar la seguridad ciudadana en América Latina y el Caribe. Amenazas, desafíos y nudos estratégicos de gestión. *Banco Interamericano de Desarrollo*, 1–47. [https://publications.iadb.org/bitstream/handle/11319/5100/Gobernar la seguridad ciudadana en América Latina y el Caribe.pdf?sequence=1](https://publications.iadb.org/bitstream/handle/11319/5100/Gobernar%20la%20seguridad%20ciudadana%20en%20América%20Latina%20y%20el%20Caribe.pdf?sequence=1)

- Ben Mabrouk, A., & Zagrouba, E. (2017). Spatio-temporal feature using optical flow based distribution for violence detection. *Pattern Recognition Letters*, 92, 62–67.
<https://doi.org/10.1016/j.patrec.2017.04.015>
- Bouachir, W., Gouiaa, R., Li, B., & Noumeir, R. (2018). Intelligent video surveillance for real-time detection of suicide attempts. *Pattern Recognition Letters*, 110, 1–7.
<https://doi.org/10.1016/j.patrec.2018.03.018>
- Carrión M., F. (2005). La inseguridad ciudadana en América Latina. *Quórum. Revista de Pensamiento Iberoamericano*, 12, 29–52. <https://www.redalyc.org/pdf/520/52001204.pdf>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(Sept. 28), 321–357. <https://dl.acm.org/doi/10.5555/1622407.1622416>
- Chebi, H., Acheli, D., & Kesraoui, M. (2016). Dynamic detection of abnormalities in video analysis of crowd behavior with DBSCAN and neural networks. *Advances in Science, Technology and Engineering Systems*, 1(5), 56–63. <https://doi.org/10.25046/aj010510>
- Chen, L. H., Hsu, H. W., Wang, L. Y., & Su, C. W. (2011). Violence detection in movies. *Proceedings - 2011 8th International Conference on Computer Graphics, Imaging and Visualization, CGIV 2011*, 119–123. <https://doi.org/10.1109/CGIV.2011.14>
- Chun, Y. D., Kim, N. C., & Jang, I. H. (2008). Content-based image retrieval using multiresolution color and texture features. *IEEE Transactions on Multimedia*, 10(6), 1073–1084. <https://doi.org/10.1109/TMM.2008.2001357>
- Clarín, C. T., M. Dionisio, J. A., Echavez, M. T., & Naval, P. C. (2005). DOVE: Detection of movie violence using motion intensity analysis on skin and blood. *Pcsc*, 6, 150–156.
https://www.researchgate.net/publication/249918692_DOVE_Detection_of_Movie_Violenc

e_using_Motion_Intensity_Analysis_on_Skin_and_Blood

- Condori Huahuachampi, L., Quispe Huamaní, G. E., Peña Carhuapoma, J. V., Carhuachin Ale, Gleen Daniel Durand Carrión, Dilicia Hilario Campos, Verónica Peña Aldazaba, Henry Cáceres Blas, J., & Cabello Herencia, V. (2018). Informe Técnico: Estadística de Seguridad Ciudadana 2018. *INEI, N°4*, 1–162. <https://www.inei.gob.pe/biblioteca-virtual/boletines/estadisticas-de-seguridad-ciudadana/1/>
- Da Silva, M. M., Evin, D. A., & Verrastro, S. (2016). Speaker-independent embedded speech recognition using Hidden Markov Models. *CACIDI 2016 - Congreso Argentino de Ciencias de La Informatica y Desarrollos de Investigacion*. <https://doi.org/10.1109/CACIDI.2016.7785985>
- Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, 512–515. <https://doi.org/10.1609/icwsm.v11i1.14955>
- De Souza, F. D. M., Chávez, G. C., Do Valle, E. A., & Araújo, A. D. A. (2010). Violence detection in video using spatio-temporal features. *Proceedings - 23rd SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2010*, 224–230. <https://doi.org/10.1109/SIBGRAPI.2010.38>
- Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior Recognition via Sparse Spatio-Temporal Features. *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 65–72. <https://doi.org/10.1109/VSPETS.2005.1570899>
- Dominguez-Sanchez, A., Cazorla, M., & Orts-Escolano, S. (2017). *Pedestrian Movement*

- Direction Recognition Using Convolutional Neural Networks*. 18(12), 3540–3548.
<https://doi.org/10.1109/TITS.2017.2726140>
- Dong, Z., Qin, J., & Wang, Y. (2016). Multi-stream deep networks for person to person violence detection in videos. *Communications in Computer and Information Science*, 662, 517–531.
https://doi.org/10.1007/978-981-10-3002-4_43
- Durand Carrión, D., Cáceres Blas, J., Peña Aldazabal, H., Hilario Campos, V., & Cabello Herencia, V. (2017). Anuario Estadístico de la Criminalidad y Seguridad Ciudadana 2011 - 2017. *INEI*, 5(9), 290.
https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1534/libro.pdf
- Farkhadov, M., Smirnov, V., & Eliseev, A. (2018). Application of speech analytics in information space monitoring systems. *2017 5th International Conference on Control, Instrumentation, and Automation, ICCIA 2017, 2018-Janua*, 92–97.
<https://doi.org/10.1109/ICCIAutom.2017.8258659>
- Fausett, L. (1994). Fundamentals of neural networks: architectures, algorithms, and applications. In *Prentice-Hall, Inc*. <https://dl.acm.org/doi/10.5555/197023>
- Fukui, H., Yamashita, T., Yamauchi, Y., Fujiyoshi, H., & Murase, H. (2015). Pedestrian detection based on deep convolutional neural network with ensemble inference network. *IEEE Intelligent Vehicles Symposium, Proceedings, 2015-Augus(Iv)*, 223–228.
<https://doi.org/10.1109/IVS.2015.7225690>
- Garmendia Lorena, F. (2011). La violencia en América Latina. *Anales de La Facultad de Medicina*, 72(4), 269–276.
http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1025-

55832011000400008&lng=es&tlng=es

- Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., & Theodoridis, S. (2006). Violence content classification using audio features. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3955 LNAI, 502–507. https://doi.org/10.1007/11752912_55
- Haykin, S. (1998). Neural Networks: A Comprehensive Foundation. In *Prentice Hall PTR*. <https://dl.acm.org/doi/book/10.5555/521706>
- Izadi, M., Azimifar, Z., & Jowkar, G. H. (2018). Abnormal event detection in indoor video using feature coding. *19th CSI International Symposium on Artificial Intelligence and Signal Processing, AISP 2017, 2018-Janua*, 151–155. <https://doi.org/10.1109/AISP.2017.8324127>
- Kahlil, M., Faris, R., Muhammad Rizky, M., Prayuda Juniarta, D. A., Richard, D., & Indra, N. (2019). A unified smart surveillance system incorporating adaptive foreground extraction and deep learning-based classification. *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*. <https://doi.org/10.1109/ICAIIIC.2019.8669017>
- Kang, K., Ouyang, W., Li, H., & Wang, X. (2016). Object detection from video tubelets with convolutional neural networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem*, 817–825. <https://doi.org/10.1109/CVPR.2016.95>
- Kiktova-Vozarikova, E., Juhar, J., & Cizmar, A. (2015). Feature selection for acoustic events detection. *Multimedia Tools and Applications*, 74(12), 4213–4233. <https://doi.org/10.1007/s11042-013-1529-2>
- Kim, C., Lee, J., Han, T., & Kim, Y. M. (2018). A hybrid framework combining background

- subtraction and deep neural networks for rapid person detection. *Journal of Big Data*, 5(1).
<https://doi.org/10.1186/s40537-018-0131-x>
- Kim, J., & Grauman, K. (2010). *Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates*. 2921–2928.
<https://doi.org/10.1109/cvpr.2009.5206569>
- Ko, T. (2008). A survey on behavior analysis in video surveillance for homeland security applications. *Proceedings - Applied Imagery Pattern Recognition Workshop*.
<https://doi.org/10.1109/AIPR.2008.4906450>
- Lai, P. K., Decombas, M., Moutet, K., & Laganier, R. (2016). Video summarization of surveillance cameras. *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2016, June 2015*, 286–294.
<https://doi.org/10.1109/AVSS.2016.7738018>
- Lai, Y. H., & Yang, C. K. (2015). Video object retrieval by trajectory and appearance. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(6), 1026–1037.
<https://doi.org/10.1109/TCSVT.2014.2358022>
- Latinobarómetro. (2017). Informe 2017. *Latinobarómetro*, 2017(Fase 1), 66.
http://www.crm.org.mx/PDF/INF_ACTIVIDADES/INFORME2016.pdf
- Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face Recognition: A Convolutional Neural-Network Approach. *IEEE Transactions on Neural Networks*, 8(1), 98–113. <https://doi.org/10.1109/72.554195>
- Lee, S., Kim, H., Lee, S., Kim, Y., Lee, D., Ju, J., & Myung, H. (2014). Detection of a Suicide by Hanging Based on a 3-D Image Analysis. *IEEE Sensors Journal*, 14(9), 2934–2935.
<https://doi.org/10.1109/JSEN.2014.2332070>

- Li, J., Liang, X., Shen, S., Xu, T., Feng, J., & Yan, S. (2018). Scale-aware Fast R-CNN for Pedestrian Detection. *IEEE Transactions on Multimedia*, 20(4), 985–996.
<https://doi.org/10.1109/TMM.2017.2759508>
- Liu, J., Zhang, S., Wang, S., & Metaxas, D. N. (2016). Multispectral deep neural networks for pedestrian detection. *British Machine Vision Conference 2016, BMVC 2016, 2016-Septe*, 73.1-73.13. <https://doi.org/10.5244/c.30.73>
- Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010). Anomaly detection in crowded scenes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1975–1981. <https://doi.org/10.1109/CVPR.2010.5539872>
- Morris, B. T., & Trivedi, M. M. (2011). Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11), 2287–2301.
<https://doi.org/10.1109/TPAMI.2011.64>
- Munezero, M., Montero, C. S., Kakkonen, T., Sutinen, E., Mozgovoy, M., & Klyuev, V. (2014). Automatic detection of antisocial behaviour in texts. *Informatica (Slovenia)*, 38(1), 3–10.
- Nilsson, J., Andersson, P., Yu-Hua Gu, I., & Fredriksson, J. (2014). Pedestrian detection using augmented training data. *Proceedings - International Conference on Pattern Recognition*.
<https://doi.org/10.1109/ICPR.2014.778>
- Noorit, N., Suvonvorn, N., & Karnchanadecha, M. (2010). Model-based human action recognition. *Second International Conference on Digital Image Processing*, 7546, 75460P.
<https://doi.org/10.1117/12.853223>
- Pathan, S. S., Al-Hamadi, A., & Michaelis, B. (2010). Crowd behavior detection by statistical modeling of motion patterns. *Proceedings of the 2010 International Conference of Soft*

- Computing and Pattern Recognition, SoCPaR 2010*, 81–86.
<https://doi.org/10.1109/SOCPAR.2010.5686403>
- Prakash Gupta, J., Dixit, P., Singh, N., & Bhaskar Semwal, V. (2014). Analysis of Gait Pattern to Recognize the Human Activities. *International Journal of Interactive Multimedia and Artificial Intelligence*, 2(7), 7. <https://doi.org/10.9781/ijimai.2014.271>
- Pritch, Y., Rav-Acha, A., & Peleg, S. (2008). Nonchronological video synopsis and indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11), 1971–1984.
<https://doi.org/10.1109/TPAMI.2008.29>
- PUCP. (2016). REPORTE DE RESULTADOS DE SONDEO NACIONAL URBANO RURAL MARZO 2016. In *Instituto de Opinión Pública*. <https://textos.pucp.edu.pe/texto/REPORTE-DE-RESULTADOS-DE-SONDEO-NACIONAL-URBANO-RURAL-MARZO-20161>
- Quispe Huamaní, G. E., Carhuachin Ale, G. D., Peña Carhuapoma, J. V., Carbajal Requena, M. A., Wilson Escribas, K. Y., de Cossio Zunino, F., & Vallejo Pereyra, R. (2018). Victimización en el Perú 2010 - 2017. In *INEI*.
https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1519/index.html
- Reyna, C., & Toche, E. (1999). La inseguridad en el Perú. *Comisión Económica Para América Latina y El Caribe*, 29, 46. <https://repositorio.cepal.org/handle/11362/6261>
- Roa, J., Jacob, G., Gallino, L., & Hung, P. C. K. (2018). Towards Smart Citizen Security Based on Speech Recognition. *Congreso Argentino de Ciencias de La Informatica y Desarrollos de Investigacion, CACIDI 2018*, 1–6. <https://doi.org/10.1109/CACIDI.2018.8584192>
- Rota, N., & Thonnat, M. (2000). Video sequence interpretation for visual surveillance. *Proceedings - 3rd IEEE International Workshop on Visual Surveillance, VS 2000*, 59–68.

<https://doi.org/10.1109/VS.2000.856858>

Schedi, M., Sjöberg, M., Mironică, I., Ionescu, B., Quang, V. L., Jiang, Y.-G., & Demarty, C.-H. (2015). VSD2014: A Dataset for Violent Scenes Detection in Hollywood Movies and Web Videos. *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*. <https://doi.org/10.1109/CBMI.2015.7153604>

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *SocialNLP 2017 - 5th International Workshop on Natural Language Processing for Social Media, Proceedings of the Workshop AFNLP SIG SocialNLP, 2012*, 1–10. <https://doi.org/10.18653/v1/w17-1101>

Shih, H. C. (2013). A novel attention-based key-frame determination method. *IEEE Transactions on Broadcasting*, 59(3), 556–562. <https://doi.org/10.1109/TBC.2013.2265782>

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., & Blake, A. (2011). Real-time human pose recognition in parts from single depth images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1297–1304. <https://doi.org/10.1109/CVPR.2011.5995316>

Stamos, D., Martelli, S., Nabi, M., McDonald, A., Murino, V., & Pontil, M. (2015). Learning with dataset bias in latent subcategory models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June*, 3650–3658. <https://doi.org/10.1109/CVPR.2015.7298988>

Sudhakaran, S., & Lanz, O. (2017). Learning to detect violent videos using convolutional long short-term memory. *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017*. <https://doi.org/10.1109/AVSS.2017.8078468>

Susarla, P., Agrawal, U., & Jayagopi, D. B. (2018). Human weapon-activity recognition in

- surveillance videos using structural-RNN. *ACM International Conference Proceeding Series*, 2018-March(June), 101–107. <https://doi.org/10.1145/3177148.3180080>
- Tejada, E. D., & Rodriguez, P. A. (2017). Moving object detection in videos using principal component pursuit and convolutional neural networks. *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 793–797. <https://doi.org/10.1109/GlobalSIP.2017.8309069>
- Thomas, S. S., Gupta, S., & Subramanian, V. K. (2017). Smart surveillance based on video summarization. *TENSYMP 2017 - IEEE International Symposium on Technologies for Smart Cities*. <https://doi.org/10.1109/TENCONSpring.2017.8070003>
- Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1). <https://doi.org/https://doi.org/10.1162/jocn.1991.3.1.71>
- Vega, I. R., & Sarkar, S. (2003). Statistical motion model based on the change of feature relationships: Human gait-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10), 1323–1328. <https://doi.org/10.1109/TPAMI.2003.1233906>
- Won, D., Steinert-Threlkeld, Z. C., & Joo, J. (2017). Protest activity detection and perceived violence estimation from social media images. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, 786–794. <https://doi.org/10.1145/3123266.3123282>
- Ya, W., Bao, T., Chunhui, D., & Zhu, M. (2017). Face recognition in real-world surveillance videos with deep learning method. *2017 2nd International Conference on Image, Vision and Computing, ICIVC 2017*, 239–243. <https://doi.org/10.1109/ICIVC.2017.7984553>
- Zhang, R., Vogler, C., & Metaxas, D. (2007). Human gait recognition at sagittal plane. *Image and Vision Computing*, 25(3), 321–330. <https://doi.org/10.1016/j.imavis.2005.10.007>
- Zhang, T., Jia, W., Gong, C., Sun, J., & Song, X. (2018). Semi-supervised dictionary learning via

local sparse constraints for violence detection. *Pattern Recognition Letters*, 107, 98–104.

<https://doi.org/10.1016/j.patrec.2017.08.021>

Zhang, W., Shan, S., Gao, W., Chen, X., & Zhang, H. (2005). Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A Novel Non-Statistical Model for Face Representation and Recognition. *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, I*, 786–791. <https://doi.org/10.1109/ICCV.2005.147>

Zhao, B., Fei-Fei, L., & Xing, E. P. (2011). Online detection of unusual events in videos via dynamic sparse coding. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3313–3320.

<https://doi.org/10.1109/CVPR.2011.5995524>

Zhou, X., Yang, C., & Yu, W. (2013). Moving object detection by detecting contiguous outliers in the low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3), 597–610. <https://doi.org/10.1109/TPAMI.2012.132>