ISSN: 1692-7257 - Volumen 1 – Número 43 - 2024

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v1i43.2888

Detección de ataques de presentación facial basado en siamese-LSTM y el análisis del flujo óptico y puntos de referencia facial

Face presentation attack detection based on siamese-LSTM and analysis of optic flow and facial landmarks

Ing. Arnold Jiménez Vargas [1], PhD. Rubiel Vargas Cañas [2]
PhD. Carlos Alberto Cobos Lozada [1], PhD. Humberto Loaiza Correa [3]

[1] ***Universidad del Cauca,*** *Facultad de Ingeniería Electrónica y Telecomunicaciones, Grupo de I+D en Tecnologías de la Información, Popayán, Cauca, Colombia.*
[2] ***Universidad del Cauca,*** *Facultad de Ciencias Naturales, Exactas y de la Educación, Grupo de Investigación en Sistemas Dinámicos, Instrumentación y Control, Popayán, Cauca, Colombia.*
[3] ***Universidad del Valle,*** *Facultad de Ingeniería, Grupo de Percepción y Sistemas Inteligentes, Santiago de Cali, Valle del Cauca, Colombia.*

Correspondencia: *rubiel@unicauca.edu.co*

Recibido: *8 noviembre 2023.* ***Aceptado:*** *10 enero 2024.* ***Publicado:*** *30 abril 2024.*

Cómo citar: *A. J. Jimenez Vargas, R. Vargas Cañas, C. A. Cobos Lozada, y H. Loaiza Correa, «Detección de ataques de presentación facial basado en siamese-LSTM y el análisis del flujo óptico y puntos de referencia facial», RCTA, vol. 1, n.º 43, pp. 125–133, abr. 2024. Recuperado de https://ojs.unipamplona.edu.co/index.php/rcta/article/view/2888*

Resumen: La autenticación por medio de la biometría facial se ha vuelto fundamental para verificar la identidad de las personas en transacciones en línea, ya que mecanismos clásicos como la autenticación por nombre de usuario y contraseña han demostrado ser poco fiables, ya que los usuarios suelen escoger contraseñas que son fáciles de recordar. Sin embargo, el avance en la fabricación de modelos con materiales como el látex, el aumento en la calidad de las impresiones y la mejora en las resoluciones de las pantallas han exigido que los sistemas de detección de fraude se adapten rápidamente a las nuevas condiciones. El presente trabajo muestra una propuesta para abordar el problema de la detección de ataques de presentación por medio de la extracción del flujo óptico y los puntos de referencia facial y su análisis por medio de una red siamese. Para evaluar el modelo propuesto, se utilizaron tres data sets: Rose-youtu, Replay-attack y Replay-mobile, y las métricas HTER y EER.

Palabras clave: Biometría, contra la suplantación de identidad, red neuronal siamese, red LSTM, flujo óptico, puntos de referencia faciales.

Abstract: Facial biometrics authentication has become essential in verifying the identity of individuals in online transactions, as classic mechanisms like username and password authentication have proven unreliable due to users often choosing easily memorable passwords. However, advances in model manufacturing with materials such as latex, print quality improvements, and screen resolution enhancements have demanded that fraud detection systems quickly adapt to new conditions. This paper proposes to address the problem of detecting presentation attacks by extracting optical flow and facial landmarks and analyzing them through a Siamese-LSTM network. The proposed model was evaluated

using three datasets: Rose-youtu, Replay-attack, and Replay-mobile, and two metrics: HTER and EER.

Keywords: Biometrics, anti-spoofing, Siamese neural network, LSTM network, optical flow, facial landmark points.

1. INTRODUCCIÓN

El uso masivo de dispositivos móviles, como las tabletas o los smartphones, ha permitido que se ofrezcan servicios por medio de aplicaciones móviles que requieren la identificación inequívoca de los usuarios. En escenarios tales como los servicios financieros u orientados a la seguridad, la autenticación por nombre de usuario y contraseña ha mostrado no ser la más indicada y es en estos contextos donde el uso de autenticación biométrica es fundamental. La autenticación biométrica se basa en la extracción y procesamiento de rasgos físicos de las personas para verificar su identidad, rasgos que pueden ser extraídos de la voz, las huellas dactilares, el rostro, entre otros. La ventaja de la biometría facial radica en que del rostro se pueden extraer más características que de la voz o de las huellas dactilares, pero debe lidiar con el problema conocido como los ataques de presentación (Presentation attack detection) [1]. Uno de los problemas más desafiantes en esta área, consiste en determinar si el dispositivo está capturando a una persona y no una fotografía o una pantalla que muestra un video o una foto, modalidades de fraude que muestran ser las más comunes en los ataques de presentación [2].

Los métodos de detección de ataques de presentación se dividen básicamente en dos: Clásicos y basados en deep learning [3]. Los métodos clásicos incluyen el análisis de movimiento por medio de técnicas como DMD (Dynamic Mode Decomposition) [4] utilizadas en otras áreas como en la dinámica de fluidos, y el análisis de textura por medio de características extraídas manualmente, como LBP (local binary pattern) [5] o mediante el operador Sobel [6]. El problema con los enfoques basados en texturas es que son sensibles a las condiciones de iluminación, a los dispositivos de captura, entre otros factores; mientras que los enfoques basados en movimiento son sensibles a ataques en los que se incluye el movimiento de ojos o labios, como en ataques de impresión o de máscaras de látex [7]. Los métodos basados deep learning, que han adquirido notoriedad en los últimos años debido al rápido avance de esta tecnología, permiten entrenar modelos capaces de extraer las características de manera autónoma, es decir, sin la supervisión humana y a través de la revisión de las muestras de los dataset, esto permite a los modelos inferir características que no son del todo perceptibles al ojo humano, y por medio de técnicas como transfer learning [8], aprender nuevas características a partir de modelos previamente entrenados.

Este trabajo aborda el problema de la detección de ataques de presentación a través de una red siamese compuesta por dos ramas de redes LSTM, las cuales son alimentadas con dos tipos de datos: el flujo óptico calculado para pares de fotogramas de videos de corta duración los cuales están separados entre sí por 300 milisegundos, y 68 puntos de referencia facial que son extraídos de cada uno de los fotogramas. Las redes siamese se han utilizado en la comparación de imágenes y consisten en dos ramas idénticas de una red neuronal que comparten los mismos pesos y están alimentadas con las dos imágenes a comparar. La salida de ambas ramas se concatena y se utiliza para determinar si las imágenes (por ejemplo, rostros o firmas) corresponden a la misma persona [9]. Esta capacidad de las redes neuronales para determinar el grado de similitud entre las dos entradas se aplicó al proceso de detección de ataques de presentación al combinarla con redes LSTM, que son redes muy utilizadas en tareas de procesamiento de datos secuenciales como las series de tiempo [10] y documentos o texto [11]. Los datos temporales están constituidos por el flujo óptico y los puntos de referencia facial, donde el flujo óptico brinda información sobre el movimiento aparente de los objetos en el video y los puntos de referencia facial aportan información sobre los movimientos de los puntos clave del rostro.

Lo que resta de este documento está organizado como se describe a continuación, en la sección 2 se presentan los trabajos previos más recientes y relevantes relacionados con la detección de fraudes en biometría facial. En la sección 3, se presentan conceptos básicos necesarios para comprender el modelo propuesto, incluyendo la arquitectura de la red siamese, la red LSTM, la red siamese-LSTM, el flujo óptico y los puntos faciales representativos. En

la sección 4 se presenta el modelo propuesto, proveyendo los detalles necesarios para comprenderlo y replicarlo. En la sección 5 se presentan la experimentación realizada, partiendo de la descripción de los dataset seleccionados para entrenar y evaluar el modelo, las métricas de comparación y los resultados obtenidos por el modelo propuesto en comparación con los reportados en el estado del arte. Finalmente, en la sección 6, se presentan las conclusiones y el trabajo futuro que se espera abordar en el tema.

2. TRABAJOS RELACIONADOS

En esta sección se revisan algunos trabajos previos en el campo de la detección de ataques de presentación, donde algunos de ellos combinan diferentes enfoques, como los orientados al análisis de texturas, de movimiento o basados en deep learning.

En [12] se encuentra un modelo basado en el flujo óptico y el análisis de texturas que extrae el flujo óptico (sección 3.4) de secuencias de vídeo para describir la dirección y la intensidad del movimiento, y luego los integra con información de textura. El método también introduce mecanismos de atención local y de canal para asignar de forma adaptativa los pesos de las distintas regiones y canales, respectivamente. En [13] los autores proponen un operador de gradiente que aprende, diseñado para extraer información eficiente de grano fino, como la magnitud del gradiente espacial, para distinguir las imágenes fraudulentas de las genuinas. Este enfoque ofrece una solución basada en datos, a diferencia de métodos como el operador Sobel que es creado a mano y utiliza pesos fijos para calcular la magnitud del gradiente de una imagen.
Otros modelos usan información del cliente como parte del proceso de detección de un ataque de presentación, por ejemplo, en [14], los autores proponen un enfoque para la detección de ataques de presentación utilizando clasificadores específicos para cada cliente y redes neuronales convolucionales. De acuerdo con los autores, los sistemas de detección de ataques de presentación obtienen la información relevante de cada clase, mientras los métodos específicos de cada cliente tienen en cuenta las características de cada individuo, ya que finalmente los sistemas de autenticación se utilizan para garantizar el acceso a usuarios de manera individual, así, con este enfoque, el sistema es capaz de adaptarse a las características y patrones únicos de cada individuo, mejorando la precisión en la detección. En [15] los autores proponen un enfoque que utiliza información sobre la identidad del cliente, reconociendo primero la cara y luego su identidad; la identidad del cliente se utiliza para seleccionar una imagen real del sujeto con la que se forma una pareja con la imagen de entrada, esta pareja se utiliza como entrada en una red Siamese que determina si es una pareja positiva (ambas imágenes son genuinas), o si la pareja es negativa (una de las imágenes es fraudulenta).

3. CONCEPTOS BÁSICOS DEL MODELO

3.1. Red neuronal siamese

La red neuronal siamese es una arquitectura de red neuronal diseñada para comparar dos o más entradas y determinar el nivel de similitud. Esta arquitectura consiste en dos o más ramas idénticas, donde cada una toma una entrada y la procesa de manera independiente, los resultados se unen en una capa de salida en la cual y basado en una métrica de distancia se calcula la similitud entre las entradas.

La Fig. 1 muestra una red siamese con dos entradas y sus componentes, que se describen a continuación.

- Input: Esta red recibe dos entradas marcadas como $Input_1$ e $Input_2$, que representan los datos que se desean comparar. Los datos deben ser procesados previamente para convertirlos a una representación usable por las redes neuronales que componen las ramas.
- Network Branch: Rama compuesta por capas de neuronas. Esta red consta de dos o más ramas idénticas, que comparten la misma arquitectura y parámetros. Se puede decir que se tiene una sola red neuronal que se replica según el número de ramas que tenga la red siamese.
- Encodings: Son las características extraídas por las redes neuronales de las ramas al procesar las entradas, por ejemplo, $Encoding_1$ y $Encoding_2$.
- Distance function: Es la capa encargada de realizar la comparación de los encodings. Aquí se utilizan funciones como la distancia euclidiana.
- Classification: Es la capa encargada de realizar la clasificación de acuerdo con el resultado entregado por la función de distancia. Está compuesta generalmente por una serie de capas densas.
- Output: La salida de la red neuronal varía de acuerdo con el problema que se está resolviendo, clasificación, regresión o una medida de similitud entre las entradas. En clasificación de imágenes

puede por ejemplo decir si dos imágenes de entrada pertenecen a la misma clase.

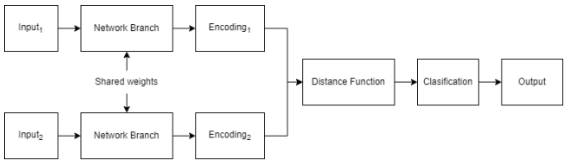

Fig. 1. *Arquitectura una red neuronal siamese*
Fuente: *(adaptada de [1])*

3.2. Long short-Term memory networks (LSTM)

Una red neuronal LSTM es un tipo de red neuronal recurrente (Recurrent Neural Networks, RNN) que está diseñada para resolver el problema del desvanecimiento del gradiente en las RNN tradicionales, el cual ocurre cuando los gradientes utilizados para actualizar los pesos en la red se vuelven muy pequeños, lo que le dificulta a la red aprender dependencias a largo plazo. Esto lo hace controlando el flujo de información a través de las compuertas.

La Fig. 2 muestra una red LSTM y sus componentes. La entrada de una red LSTM consiste en:

- Input (X_t): Vector de entrada en el tiempo t. La dimensionalidad de este vector depende del dominio del problema y de la tarea específica a realizar.
- Previous hidden state (H_{t-1}): Salida del paso anterior t-1. El estado oculto es un vector que resume la memoria de la red respecto a las entradas anteriores.
- Previous cell state (C_{t-1}): la celda de estado también es una salida del paso anterior t-1. Representa el estado interno de la célula, se comporta como una forma de memoria.

Las compuertas (gates) en la LSTM se usan para aprender y olvidar información de manera selectiva a través del tiempo, lo que les permite retener información sobre secuencias de datos más largas. A continuación, se presentan los tipos de compuertas principales:

- Forget gate (F_t): Determina qué información del estado oculto anterior y la entrada actual deben olvidarse. Su salida es un valor entre 0 y 1 por cada elemento del estado oculto, donde cero se utiliza para <<olvidar totalmente >> y 1 para <<mantener o recordar totalmente>>.
- Input gate (I_t) or Update gate: Determina qué información nueva de la entrada actual debería añadirse al estado oculto. Su salida es un valor entre 0 y 1 por cada elemento del estado oculto, donde 0 significa <<ignorar por completo la nueva información>> y 1 significa <<agregar completamente la nueva información>>.
- Output gate (O_t): Determina qué información del estado oculto actual debería salir. Su salida es un valor entre 0 y 1 por cada elemento del estado oculto, donde 0 significa <<ignorar completamente la información>> y 1 significa <<emitir completamente la información>>.

La Fig. 3. muestra una red LSTM desenrollada para una entrada de tres elementos en los tiempos t, t+1 y t+2. La entrada a la celda en el tiempo t consiste solo en el valor X_t, mientras que para la segunda es el valor X_{t+1} y los vectores H_t y C_t, esto porque el estado de la memoria es inexistente en el tiempo inicial t. La salida de la celda LSTM en cada momento H_t, constituye los encodings que luego pueden ser utilizados, por ejemplo, en una red densa para otras tareas.

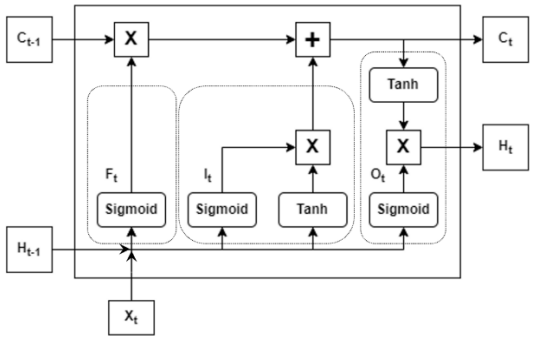

Fig. 2. *Arquitectura una red neuronal LSTM*
Fuente: *(adaptada de [2])*

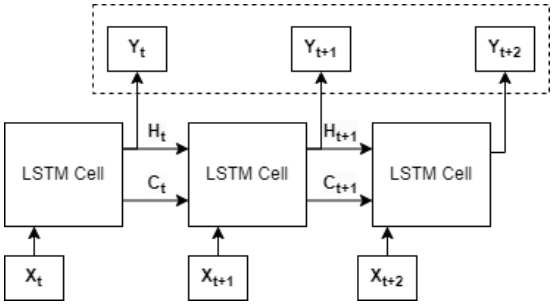

Fig. 3. *Red LSTM para una entrada de tres elementos*
Fuente: *(adaptada de [3])*

3.3. Red neuronal siamese-LSTM

En una red Siamese-LSTM las ramas están constituidas por redes LSTM idénticas, ya que comparten los mismos parámetros según se mostró en la sección 3.1. Estas redes son útiles cuando se quiere comparar pares de datos de naturaleza temporal, tales como series de tiempo o videos. La Fig. 4 muestra la arquitectura de una red Siamese-LSTM, donde X_1 y X_2 corresponden a los vectores

de entradas de naturaleza secuencial, pudiendo ser frases, por ejemplo. El componente LSTM Cell es una representación simplificada de la red neuronal LSTM en la que se ha representado un bucle de retroalimentación indicando que parte de la entrada de una celda LSTM la constituye la salida en el tiempo anterior de la celda LSTM, cada rama entrega los encodings representados por Y_1 y Y_2 que pasan por una función de distancia que calcula la similitud que finalmente pasa, en este caso, por una red densa encargada de realizar la clasificación.

3.4. Flujo óptico

El flujo óptico es una técnica utilizada para determinar el patrón de movimiento aparente de los objetos entre dos imágenes o secuencia de vídeo, causado por el movimiento de la cámara o de los objetos de la escena. Se utiliza en el contexto del procesamiento de video para estimar el movimiento de objetos en cuadros consecutivos, analizando los cambios en la intensidad y los colores de los pixeles en los cuadros a través del tiempo. Los algoritmos de flujo óptico pueden estimar el desplazamiento de objetos en una escena y crear un campo vectorial que describe la dirección y la magnitud del movimiento en cada punto del cuadro. La Fig. 5 muestra el modelo RAFT [5] (Recurrent All Pairs Field Transforms for Optical Flow) que a partir de dos imágenes (parte superior izquierda) calcula el flujo óptico (parte inferior derecha), el cual fue usado en la presente investigación.

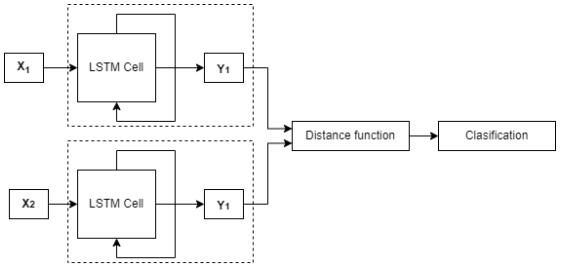

Fig. 4. *Arquitectura de red Siamese-LSTM*
Fuente: *(adaptada de [4])*

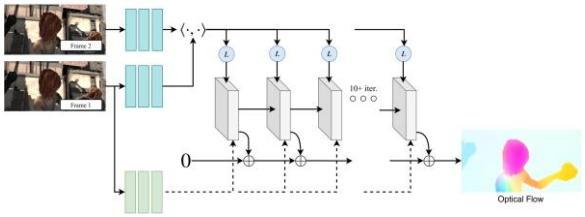

Fig. 5. *Arquitectura del modelo RAFT*
Fuente: *(tomada de [5])*

3.5. Facial landmark points

Los puntos de referencia facial (facial landmark points) son puntos o rasgos específicos de la cara que pueden utilizarse para identificar y medir distintas características faciales. Los puntos de referencia facial incluyen: la punta de la nariz, los bordes exteriores de los ojos, las comisuras de los labios, las esquinas de la mandíbula y el puente de la nariz y las cejas. Estos puntos se utilizan para medir la distancia entre diferentes áreas de la cara, la angulación de los rasgos faciales, la relación entre diferentes partes de la cara y la simetría facial. La Fig. 6 muestra los puntos de referencia facial calculados para un fotograma de uno de los videos del dataset Rose-youtu. Generalmente se trabaja con 68 puntos faciales, aunque también existe la opción de usar 106 puntos.

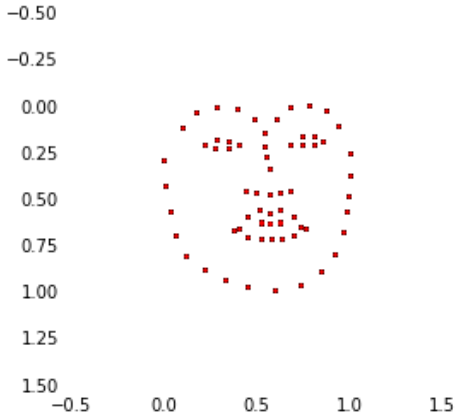

Fig. 6. *Puntos de referencia facial*
Fuente: *elaboración propia.*

4. MODELO PROPUESTO

El modelo propuesto para la detección de ataques de presentación en la autenticación por biometría facial consiste en una red Siamese-LSTM alimentada con registros conformados por información del flujo óptico y los puntos de referencia facial. La Fig. 7 muestra el proceso general, que consiste en la extracción de los fotogramas con los que se obtienen los puntos de referencia facial y el flujo óptico, los cuales se combinan para formar los registros de entrada. Después, con los registros generados se realiza la ejecución de la red Siamese-LSTM para obtener los encodings y finalmente se realiza el proceso de clasificación, donde se indica si los dos videos corresponden a la misma clase (genuino o ataque).

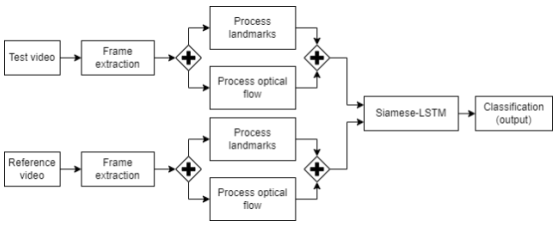

Fig. 7. *Proceso general.*
Fuente: *elaboración propia.*

De cada video se toman 30 frames separados por 300 milisegundos, a cada uno de estos frames se le calculan los puntos de referencia facial y a cada par el flujo óptico. Luego, los puntos de referencia facial y el flujo óptico se concatenan en un vector de dimensión 30 x 5544 (los detalles de la forma como se definen estos valores se muestran más abajo). Dado que para el primer fotograma no existe un frame anterior para calcular el flujo óptico, el primer elemento del registro está conformado por los puntos de referencia facial del primer fotograma y un vector de ceros. El elemento i-ésimo está conformado por los puntos de referencia facial del i-ésimo fotograma y el flujo óptico entre el i-ésimo fotograma y el fotograma anterior.

Los 68 puntos de referencia facial se calculan utilizando el modelo SBR [6] que recibe como entrada una imagen y genera un vector de dimensión 2x68 que se aplana en uno de dimensión 1x136.
El flujo óptico se calcula utilizando RAFT. Este recibe como entrada dos imágenes f_i y f_{i-1} del mismo tamaño, que son el i-ésimo fotograma y el fotograma anterior, y se genera un vector de dimensión 52x52x2 el cual se aplana en un vector de dimensión 1x5408.

Para realizar el proceso de entrenamiento del modelo, se generó una lista con pares de videos seleccionados así: Se recorre todo el dataset, y si el índice del i-ésimo video (muestra), es impar, se selecciona de manera aleatoria un video de la misma clase (genuino o fraudulento) de entre todo el dataset, si el índice es par, la selección se hace entre todos los registros del mismo sujeto. Con estos videos se generan dos pares, uno de ellos conformado por la muestra del dataset y un video de referencia de la misma clase y el otro por la muestra y un video de la clase opuesta. El par constituido por muestras de la misma clase se etiquetan con 1 y el otro par de muestras de diferente clase con 0. Al final el dataset de entrenamiento que contiene N muestras genera un dataset de tamaño 2*N. La Fig. 8 muestra cómo a partir de una muestra del dataset (G-Sample0) se generan dos pares: uno conformado por G-Sample0 y una muestra aleatoria del dataset G-SampleR con etiqueta 1, y el otro conformado por G-Sample0 y una muestra de la clase contraria S-SampleR con etiqueta 0.

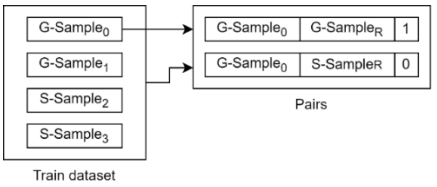

Fig. 8. *Selección de pares.*
Fuente: *elaboración propia.*

5. EXPERIMENTACIÓN

5.1. Data sets

Para evaluar la efectividad del modelo propuesto, se realizaron experimentos con los siguientes data sets:

- Rose-youtu: Este dataset [7] es una colección de videos para la detección de actividad (liveness detection) creada en la Universidad Tecnológica de Nanyan. Abarca una gran variedad de condiciones de iluminación, modelos de cámara y tipos de ataque. Consta de 4225 vídeos de 25 personas y por cada una hay entre 150 y 200 clips de vídeo, con una duración de 10 segundos en promedio. La información fue recolectada utilizando cinco dispositivos móviles diferentes: Hasee, Huawei, iPad 4, iPhone 5s y ZTE, todos ellos con cámaras frontales y una distancia de 30-50 cm entre la cara y la cámara. Los ataques de suplantación se realizaron con papel impreso, ataques de repetición de video y ataques de enmascaramiento.
- Replay-mobile: Este dataset [8] es una colección de 1190 videos de fotos y ataques de presentación de 40 personas bajo diferentes condiciones de iluminación (controlada, adversa, directa, lateral y difusa), los cuales fueron grabados utilizando un Ipad mini2 y un smartphone LG-G4. El dataset de divide en 4 grupos: Entrenamiento, desarrollo, pruebas y registro.
- Replay-attack: Este dataset [9] es una colección de 1300 videos de fotos y ataques de presentación de 50 personas bajo diferentes condiciones de iluminación. Se divide en 4 grupos: Entrenamiento, desarrollo, pruebas y registro. Los videos fueron grabados con un equipo Macbook en formato mov.

5.2. Protocolo de experimentación

En la fase de experimentación, se utilizó el conjunto de prueba de los data sets replay-attack y replay-mobile, sin embargo, el dataset rose-youtu no cuenta con un conjunto de prueba separado y, para abordar esta limitación, se generó el conjunto de prueba con los clientes 13 al 18 y 20 al 23 del dataset. Para cada registro en los conjuntos de prueba seleccionados, se formaron parejas al azar, asignando la etiqueta 1 si los pares correspondían a la misma clase y una etiqueta de 0 si pertenecían a clases diferentes. Se garantizó que por cada registro siempre se generara una pareja positiva y una negativa.

5.3. Métricas

A continuación, se describen las métricas utilizadas para evaluar la efectividad del modelo frente a otros enfoques reportados en la literatura:

- EER (Equal error rate) [10]: Es una métrica muy utilizada para evaluar sistemas biométricos, así como otros sistemas de clasificación. Es el punto de la curva ROC (Receiver Operating Characteristic) [11] en el que la Tasa de Falsos Aceptados (FAR) es igual a la Tasa de Falsos Rechazados (FRR). La curva ROC es un gráfico que muestra la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) en diferentes umbrales de clasificación. La TPR es el porcentaje de verdaderos positivos clasificados correctamente como positivos, mientras que la FPR es el porcentaje de falsos positivos clasificados incorrectamente como positivos.
- HTER (Half Total Error Rate): También conocido como Average Classification Error Rate, se define como la media de la tasa de falsos aceptados (FAR) y la tasa de falsos rechazados (FRR), donde FAR es la proporción de instancias negativas que se clasifican como positivas y FRR es la proporción de instancias positivas que se clasifican como negativas [16].

5.4. Resultados

La Tabla 1 muestra para cada dataset el resultado de EER obtenido por el modelo propuesto y los reportados en la literatura. Además, la Tabla 2 muestra los resultados de HTER.

Tabla 1: *Comparación de resultados utilizando EER como métrica*

Dataset	Método	EER%
Replay-attack	Trabajo en [13]	2.72
	Trabajo en [17]	1.26
	Trabajo en [18]	4.70
	Propuesto	9.15

Fuente: *elaboración propia*

Tabla 2: *Comparación de resultados utilizando HTER como métrica*

Dataset	Método	HTER%
Rose-youtu	Trabajo en [14]	8.13
	WA(GA+MMS+PS) [19]	5.12
	Propuesto	13.24
Replay-mobile	Localised MKL [20]	5.60
	Trabajo en [21]	8.58
	Propuesto	6.70
Replay-attack	Trabajo en [21]	0.00
	Propuesto	3.75

Fuente: *elaboración propia*

5.5. Discusión

Se observa en la Tabla 1 y en la Tabla 2 que el modelo propuesto no presenta ventajas competitivas en comparación con otros modelos del estado del arte. En la Tabla 1, se puede observar que este enfoque se posiciona en último lugar entre los 4 trabajos que utilizan el mismo conjunto de datos Replay-attack, con una diferencia de 7.89 puntos porcentuales respecto al mejor modelo y una diferencia de 4.45 puntos porcentuales respecto al modelo que le sigue. En la Tabla 2, se puede apreciar que la diferencia con el mejor modelo es de 8.12 puntos porcentuales para el conjunto de datos Rose-youtu, de 3.75 puntos porcentuales para el conjunto de datos Replay-attack, y de 1.1 puntos porcentuales para el conjunto de datos Replay-mobile. Estos resultados indican que el modelo propuesto no logra superar a los modelos existentes en los diferentes conjuntos de datos evaluados.

6. CONCLUSIÓN Y TRABAJO FUTURO

Las redes Siamese han mostrado excelentes resultados en escenarios en los que se requiere comparar dos entradas y determinar su grado de similitud, como los sistemas de validación de la firma o de la identidad de un individuo, en estos casos una imagen de entrada es comparada con una bien conocida, esto es, con una que representa una muestra genuina. Generalmente este tipo de redes

son entrenadas y probadas con parejas positivas, que consisten en dos muestras genuinas, y con parejas negativas, que consisten en una muestra genuina y la otra fraudulenta. En el presente trabajo se experimentó generando dos parejas por cada video en la que las parejas positivas estaban formadas cada una por un registro de la misma clase, pudiendo ser genuino o fraudulento, y las parejas negativas estaban formadas cada una por un registro genuino y uno fraudulento, esto permitió experimentar con la capacidad de este tipo de arquitecturas para determinar el grado de similitud entre dos registros fraudulentos, que en términos de un sistema de biometría en producción equivaldría a, dado un video de entrada, seleccionar aleatoriamente un registro de entre los del dataset de pruebas, que podría ser genuino o fraudulento y con estos registros determinar si la entrada es genuina o no, de acuerdo con la etiqueta del registro seleccionado aleatoriamente, pero los resultados muestran que no es un enfoque apropiado, puesto que al momento de realizar el registro de un cliente lo que se guarda con base de datos es un registro genuino, en el caso de que el registro de dicho cliente se haga en un ambiente controlado. Por otro lado, la necesidad de contar con registros bien conocidos que puedan ser tomados como referencia al hacer la comprobación de un video, tal y como se vio en trabajos como los presentados en [14] y en [15], en los cuales los autores muestran la utilidad de contar con información específica del cliente, y dado que finalmente los sistemas de autenticación por biometría facial son utilizados para clientes específicos, registrados previamente en el sistema, tiene sentido que el entrenamiento y las pruebas se hagan con información sobre el cliente, también se puede trabajar en estrategias para la selección de parejas, como puede verse en [22]. Al combinar la arquitectura de red Siamese con LSTM se procuró construir un modelo centrado en el análisis de información dinámica de los sujetos y su entorno por medio de los puntos de referencia facial, que identifican puntos clave del rostro y el flujo óptico que describe el movimiento aparente de los objetos de la escena y de la cámara, pero de acuerdo con los resultados, se requiere trabajar en una representación de los registros de entrada que refleje mejor la relación entre los puntos clave del rostro y el movimiento del entorno, ya que es importante obtener un modelo que a partir de los datos identifique patrones en la forma en que puntos de referencia facial cambian entre frames en un rostro genuino y cómo el movimiento del entorno se relaciona con el rostro en cuestión. Así mismo se requiere que la representación tenga en cuenta los movimientos faciales que pueden verse influenciados por las características propias de cada individuo, como el parpadeo excesivo, parálisis faciales parciales, entre otras.

RECONOCIMIENTOS

El trabajo presentado en este artículo fue parcialmente apoyado por el Grupo de Investigación y Desarrollo en Tecnologías de la Información (GTI) de la Universidad del Cauca.

REFERENCIAS

[1] S. Jia, G. Guo, Z. Xu, and Q. Wang, “Face presentation attack detection in mobile scenarios: A comprehensive evaluation,” Image and Vision Computing, vol. 93, p. 103826, Jan. 2020, doi: 10.1016/j.imavis.2019.11.004.

[2] S. Kumar, S. Singh, and J. Kumar, A Comparative Study on Face Spoofing Attacks. 2017.

[3] Y. Xin et al., “A survey of liveness detection methods for face biometric systems,” Sensor Review, vol. 37, no. 3, pp. 346–356, Jul. 2017, doi: 10.1108/SR-08-2015-0136.

[4] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, “On Dynamic Mode Decomposition: Theory and Applications,” Journal of Computational Dynamics, vol. 1, no. 2, pp. 391–421, Dec. 2014, doi: 10.3934/jcd.2014.1.391.

[5] L. Li, X. Feng, Z. Xia, X. Jiang, and A. Hadid, “Face spoofing detection with local binary pattern network,” Journal of Visual Communication and Image Representation, vol. 54, pp. 182–192, Jul. 2018, doi: 10.1016/j.jvcir.2018.05.009.

[6] Z. Wang et al., “Deep Spatial Gradient and Temporal Depth Learning for Face Anti-spoofing.” arXiv, Mar. 18, 2020. doi: 10.48550/arXiv.2003.08061.

[7] X. Tu et al., “Learning Generalizable and Identity-Discriminative Representations for Face Anti-Spoofing,” arXiv:1901.05602 [cs], Jan. 2019, Accessed: Nov. 10, 2019. [Online]. Available: http://arxiv.org/abs/1901.05602

[8] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A Survey on Deep Transfer Learning,” in Artificial Neural Networks and Machine Learning – ICANN 2018, V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, Eds., Cham:

Springer International Publishing, 2018, pp. 270–279.

[9] V. Ruiz, I. Linares, A. Sanchez, and J. F. Velez, "Off-line handwritten signature verification using compositional synthetic generation of signatures and Siamese Neural Networks," Neurocomputing, vol. 374, pp. 30–41, Jan. 2020, doi: 10.1016/j.neucom.2019.09.041.

[10] A. Niknam, H. K. Zare, H. Hosseininasab, and A. Mostafaeipour, "Developing an LSTM model to forecast the monthly water consumption according to the effects of the climatic factors in Yazd, Iran," Journal of Engineering Research, vol. 11, no. 1, p. 100028, Mar. 2023, doi: 10.1016/j.jer.2023.100028.

[11] A. Al Hamoud, A. Hoenig, and K. Roy, "Sentence subjectivity analysis of a political and ideological debate dataset using LSTM and BiLSTM with attention and GRU models," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, Part A, pp. 7974–7987, Nov. 2022, doi: 10.1016/j.jksuci.2022.07.014.

[12] L. Li, Z. Xia, J. Wu, L. Yang, and H. Han, "Face presentation attack detection based on optical flow and texture analysis," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 4, pp. 1455–1467, Apr. 2022, doi: 10.1016/j.jksuci.2022.02.019.

[13] C. Wang, B. Yu, and J. Zhou, "A Learnable Gradient operator for face presentation attack detection," Pattern Recognition, vol. 135, p. 109146, Mar. 2023, doi: 10.1016/j.patcog.2022.109146.

[14] S. Fatemifar, S. R. Arashloo, M. Awais, and J. Kittler, "Client-specific anomaly detection for face presentation attack detection," Pattern Recognition, vol. 112, p. 107696, Apr. 2021, doi: 10.1016/j.patcog.2020.107696.

[15] M. Pei, B. Yan, H. Hao, and M. Zhao, "Person-Specific Face Spoofing Detection Based on a Siamese Network," Pattern Recognition, vol. 135, p. 109148, Mar. 2023, doi: 10.1016/j.patcog.2022.109148.

[16] C. Yuan, Q. Cui, X. Sun, Q. M. J. Wu, and S. Wu, "Chapter Five - Fingerprint liveness detection using an improved CNN with the spatial pyramid pooling structure," in Advances in Computers, A. R. Hurson and S. Wu, Eds., Elsevier, 2021, pp. 157–193. doi: 10.1016/bs.adcom.2020.10.002.

[17] X. Cheng, J. Zhou, X. Zhao, H. Wang, and Y. Li, "A presentation attack detection network based on dynamic convolution and multi-level feature fusion with security and reliability," Future Generation Computer Systems, Apr. 2023, doi: 10.1016/j.future.2023.04.012.

[18] X. Shu, X. Li, X. Zuo, D. Xu, and J. Shi, "Face spoofing detection based on multi-scale color inversion dual-stream convolutional neural network," Expert Systems with Applications, vol. 224, p. 119988, Aug. 2023, doi: 10.1016/j.eswa.2023.119988.

[19] S. Fatemifar, S. Asadi, M. Awais, A. Akbari, and J. Kittler, "Face spoofing detection ensemble via multistage optimisation and pruning," Pattern Recognition Letters, vol. 158, pp. 1–8, Jun. 2022, doi: 10.1016/j.patrec.2022.04.006.

[20] S. R. Arashloo, "Unknown Face Presentation Attack Detection via Localised Learning of Multiple Kernels." 2022.

[21] "How do Siamese Networks Work in Image Recognition? | Baeldung on Computer Science." https://www.baeldung.com/cs/siamese-networks (accessed Apr. 09, 2023).

[22] G. He, F. Li, Q. Wang, Z. Bai, and Y. Xu, "A hierarchical sampling based triplet network for fine-grained image classification," Pattern Recognition, vol. 115, p. 107889, Jul. 2021, doi: 10.1016/j.patcog.2021.107889.