ReTEC

Revista Electrónica de Tecnología, Educación y Ciencia
ISSN: 2953-5654
http://retec.unsa.edu.ar
Universidad Nacional de Salta

Hashing perceptual para la integridad de imágenes de documentos: un procedimiento experimental

Sergio Rocabado[1] [2], Jorge Ramírez[1]

[1] Departamento de Informática - Facultad de Ciencias Exactas – Universidad Nacional de Salta
[2] CIUNSa – Consejo de Investigación de la Universidad Nacional de Salta
srocabado@di.unsa.edu.ar, jramirez@di.unsa.edu.ar

Revista Electrónica de Tecnología, Educación y Ciencia,

Volumen 1, Número 3, pág. 9-20, jun, 2026. ISSN: 2953-5654

Disponible en http://retec.unsa.edu.ar/

Hashing perceptual para la integridad de imágenes de documentos: un procedimiento experimental

Sergio Rocabado[1 2], Jorge Ramírez[1]

[1] Departamento de Informática - Facultad de Ciencias Exactas – Universidad Nacional de Salta
[2] CIUNSa – Consejo de Investigación de la Universidad Nacional de Salta
srocabado@di.unsa.edu.ar, jramirez@di.unsa.edu.ar

Resumen: En este trabajo se propone un procedimiento experimental reproducible orientado a la evaluación empírica de algoritmos de hashing perceptual aplicados a la verificación de la integridad de imágenes que contienen documentos textuales. El enfoque adoptado se basa en simulación mediante la generación automática de imágenes sintéticas y de variantes con modificaciones controladas, tanto semánticas como no semánticas, lo que permite superar las limitaciones asociadas al uso de conjuntos de datos reales preexistentes.
Dentro de este esquema experimental se analizan diferentes funciones de hashing perceptual, evaluando su comportamiento mediante métricas basadas en la distancia de Hamming. Los resultados obtenidos permiten caracterizar el desempeño de estos algoritmos, evidenciando dificultades para detectar alteraciones semánticas mínimas y una sensibilidad considerable frente a transformaciones globales de luminancia.
A partir de estas observaciones, el procedimiento experimental propuesto se plantea como una base metodológica que facilita la realización de estudios comparables y extensibles en escenarios similares.

Abstract: In this work, a reproducible experimental procedure is proposed for the empirical evaluation of perceptual hashing algorithms applied to the verification of the integrity of images containing textual documents. The adopted approach is based on simulation through the automatic generation of synthetic images and controlled variants incorporating both semantic and non-semantic modifications, which helps overcome the limitations associated with the use of pre-existing real-world datasets.
Within this experimental setup, several perceptual hashing functions are analyzed by evaluating their behavior using metrics based on the Hamming distance. The results obtained allow the characterization of the performance of these algorithms, revealing difficulties in detecting minimal semantic alterations as well as a considerable sensitivity to global luminance transformations.
Based on these observations, the proposed experimental procedure is intended to serve as a methodological basis that facilitates the development of comparable and extensible studies in similar scenarios.

Palabras Claves: Hashes perceptuales; integridad de documentos; imágenes de documentos textuales; evaluación empírica; distancia de Hamming**.**

1. Introducción

La creciente digitalización de documentos ha puesto de relieve la necesidad de contar con mecanismos eficaces para verificar su integridad, especialmente cuando estos documentos se gestionan en forma de imágenes que contienen información textual. En este contexto, los hashes perceptuales han sido propuestos como una alternativa para la verificación de integridad en imágenes, ya que permiten generar descriptores compactos que conservan características visuales relevantes del contenido. A diferencia de las funciones de hash criptográficas

tradicionales, estos métodos buscan producir valores similares cuando la imagen sufre transformaciones que no modifican su apariencia global, tales como cambios de formato, compresión o variaciones de brillo.

No obstante, el uso de funciones de hash para el control de integridad de activos digitales exige que dichas funciones satisfagan determinadas propiedades cuya verificación resulta, en muchos casos, compleja de establecer únicamente mediante análisis teóricos. La experiencia acumulada en el estudio de funciones hash demuestra que características consideradas seguras durante largos períodos pueden verse comprometidas a partir de avances teóricos o del incremento en la capacidad de cómputo disponible.

Un ejemplo representativo es la función MD5, propuesta por Rivest en 1991, para la cual transcurrió más de una década hasta que Wang y Yu demostraron la posibilidad de generar colisiones en tiempos computacionales viables [1]. Este antecedente ilustra las dificultades de evaluar completamente este tipo de funciones únicamente a partir de su definición formal o del algoritmo que las implementa.

En este sentido, los enfoques empíricos basados en simulación (in silico) pueden contribuir a una mejor comprensión del comportamiento de estas funciones, tal como se evidencia en diversos trabajos previos [2,3], ya que permiten observar su desempeño en escenarios controlados generados mediante software. Asimismo, la reproducibilidad constituye un requisito fundamental en este tipo de estudios, puesto que posibilita verificar, contrastar o ampliar los resultados mediante la repetición del procedimiento experimental bajo las mismas condiciones metodológicas.

En el caso particular de las funciones de hash aplicadas a la preservación de la integridad de imágenes de documentos textuales, una evaluación empírica puede aportar información relevante sobre su comportamiento y facilitar la comparación entre diferentes propuestas. Sin embargo, una limitación frecuente es la escasez de conjuntos de datos públicos que incluyan imágenes originales y manipuladas [4]. Frente a esta situación, el presente trabajo propone la generación automática de un conjunto de imágenes sintéticas, aprovechando las capacidades actuales de los modelos de lenguaje a gran escala (LLMs) para producir textos similares a los redactados por personas [5]. Estos textos pueden representarse como imágenes y modificarse de manera controlada para distintos fines experimentales.

El presente estudio presenta una descripción metodológica detallada del procedimiento experimental propuesto y un análisis empírico del comportamiento de diversas funciones de hashing perceptual realizado sobre un conjunto de imágenes sintéticas.

2. Estudios empíricos e in silico sobre hashing en criptografía

2.1. Integridad de imágenes de documentos textuales

Los mecanismos tradicionales de verificación de integridad de documentos digitales, basados en funciones de hash criptográficas, no resultan adecuados cuando el objeto de análisis es una imagen. Mientras que estos métodos permiten detectar cualquier modificación a nivel de archivo, las imágenes digitales pueden experimentar transformaciones que no alteran el contenido semántico representado, tales como cambios de formato, compresión o ajustes de brillo.

En este contexto, Swaminathan y colaboradores [6] introdujeron el concepto de hashing robusto, definiendo una función de hash como robusta cuando, utilizando una misma clave, dos imágenes visualmente similares producen valores de hash cercanos entre sí según una determinada métrica de distancia. Este enfoque permite tolerar modificaciones que no afectan la percepción visual global de la imagen.

A partir de este trabajo surgieron diversas propuestas orientadas a preservar dicha robustez [7–9]. Algunas de ellas se enfocan en identificar imágenes similares a pesar de pequeñas diferencias, lo cual resulta útil en aplicaciones como la protección de derechos de autor o la detección de copias de material gráfico. Otras propuestas se orientan a preservar el contenido de una imagen frente a modificaciones accidentales, sin considerar alteraciones introducidas por procesos habituales como la conversión de formato o la variación de parámetros visuales.

2.2. Estudios empíricos e in silico

Las funciones de hash, tanto criptográficas como perceptuales, suelen involucrar múltiples operaciones en su proceso de cómputo, lo que dificulta determinar analíticamente sus propiedades y su comportamiento frente a modificaciones específicas de la entrada. Debido a esta complejidad, una alternativa para evaluar si una función hash satisface determinados atributos consiste en analizar su comportamiento mediante estudios empíricos [10].

No obstante, este tipo de enfoques enfrenta desafíos prácticos importantes, entre los cuales se destaca la disponibilidad de conjuntos de datos adecuados al dominio en el que se desea evaluar una determinada característica. Aunque existen colecciones masivas de imágenes disponibles para investigación o entrenamiento de modelos, estas suelen estar orientadas a aplicaciones de análisis forense y no necesariamente resultan apropiadas para el estudio de la integridad semántica de documentos textuales [11–13].

Por esta razón, el presente trabajo adopta un enfoque empírico basado en simulación (in silico), mediante la generación automática y controlada de imágenes de documentos textuales, así como de variantes que incorporan alteraciones mínimas tanto en el contenido textual como en aspectos que no afectan su significado. Este enfoque permite analizar sistemáticamente el comportamiento de distintos algoritmos de hashing frente a modificaciones controladas bajo condiciones reproducibles.

3. Trabajos relacionados

En los últimos años, diversos estudios han señalado la existencia de la denominada crisis de reproducibilidad —también conocida como crisis de replicabilidad— en distintas áreas de la investigación científica [14,15]. Este fenómeno ha sido atribuido a múltiples factores [14,16], entre los que se destacan las restricciones asociadas al secreto industrial, la protección mediante patentes y el uso de software bajo licencias privativas, que limitan el acceso a datos, herramientas y procedimientos experimentales [17–20]. Como consecuencia, estas condiciones dificultan la verificación independiente de resultados y la comparación sistemática entre diferentes estudios.

Por otra parte, el hashing perceptual surgió como respuesta a la necesidad de determinar si dos imágenes digitales representan esencialmente el mismo contenido. Este tipo de técnicas se utiliza tanto para preservar la integridad del significado de una imagen como para identificar copias modificadas, descartando cambios accidentales o transformaciones que no afectan el

contenido semántico. En general, estas funciones han sido diseñadas y evaluadas principalmente en aplicaciones relacionadas con imágenes de escenas naturales, tal como se describe en [7,9], mientras que su aplicación al análisis de imágenes de documentos textuales escaneados ha recibido comparativamente menos atención en la literatura [21].

Desde esta perspectiva, la detección de manipulaciones en imágenes de documentos textuales ha seguido, en términos generales, dos enfoques principales. Por un lado, se encuentran las técnicas activas, que incluyen el uso de funciones de hashing o marcas de agua digitales. Estos métodos requieren información adicional o un procesamiento previo del contenido para permitir la verificación posterior de su integridad [22–24]. Por otro lado, existen técnicas pasivas orientadas a la detección de manipulaciones sin necesidad de información auxiliar, generalmente desarrolladas en el ámbito de la investigación forense digital [21,25].

A partir de los antecedentes revisados se observa que, si bien existen numerosos estudios empíricos sobre funciones hash y diversas propuestas de hashing perceptual, su aplicación sistemática al análisis de la integridad de imágenes de documentos textuales ha sido relativamente limitada. En particular, se identifican carencias en relación con metodologías reproducibles y con conjuntos de datos específicamente diseñados para evaluar modificaciones semánticas y no semánticas en este tipo de imágenes. En este contexto, el presente trabajo propone un procedimiento experimental basado en la generación automática de imágenes de documentos textuales y sus variantes, con el propósito de facilitar la evaluación empírica y comparativa de funciones de hashing perceptual.

4. Procedimiento experimental y herramientas

El presente trabajo adopta un procedimiento experimental orientado a la evaluación empírica de funciones de hashing perceptual aplicadas a imágenes de documentos textuales. El objetivo principal de este enfoque es permitir el análisis sistemático del comportamiento de distintos algoritmos de hashing sobre un conjunto significativo de imágenes, así como facilitar la reproducción de los experimentos por parte de otros investigadores y la comparación de resultados entre estudios independientes.

Para llevar a cabo este análisis resulta necesario disponer de un conjunto amplio de imágenes de documentos textuales junto con variantes que incorporen modificaciones controladas. A partir de estas imágenes se calculan los valores de hash correspondientes y posteriormente se analizan las diferencias entre los resúmenes generados, lo que permite estudiar la respuesta de las funciones frente a distintos tipos de alteraciones.

La obtención manual de este tipo de conjuntos de datos presenta diversas dificultades prácticas. Entre ellas se encuentran las restricciones vinculadas a derechos de autor, el tiempo requerido para digitalizar documentos mediante escaneo y la necesidad de aplicar posteriormente distintas transformaciones sobre las imágenes. Con el propósito de superar estas limitaciones, el enfoque adoptado se basa en la generación automática de las imágenes necesarias para la evaluación.

La Figura 1 ilustra el esquema general del procedimiento utilizado para la generación de variantes controladas de imágenes y el posterior análisis del comportamiento de los algoritmos de hashing perceptual.

En una primera etapa se generan imágenes que contienen contenido textual, junto con tres variantes de cada una de ellas. La primera introduce una modificación mínima destinada a alterar

el significado del texto, mientras que las otras dos corresponden a transformaciones que no afectan su contenido semántico: una versión comprimida con pérdida de información utilizando el formato JPEG y otra en la que se reduce de forma significativa el brillo de la imagen. Estas variantes permiten construir un conjunto de imágenes que presentan diferencias respecto de la original, aunque únicamente una de ellas introduce cambios semánticos.

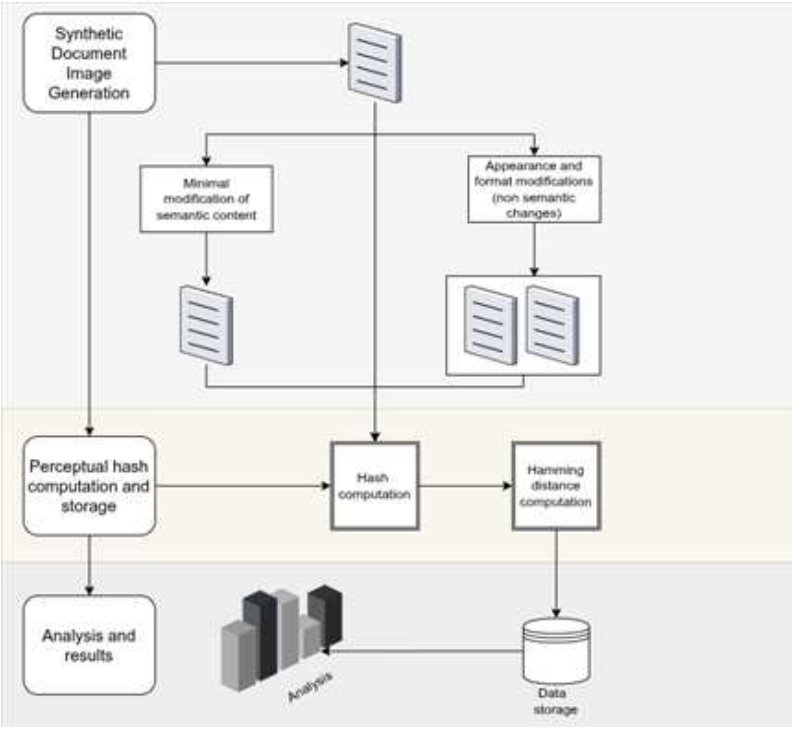

Figura 1. Esquema del procedimiento experimental para evaluar hashing perceptual en imágenes de documentos textuales.

Posteriormente se calculan los valores de hash correspondientes a cada imagen y a sus variantes. Para ello se utilizan implementaciones de funciones de hashing disponibles en bibliotecas ampliamente difundidas y probadas, distribuidas bajo licencias que permiten su uso sin restricciones en contextos académicos y de investigación.

A partir de los hashes obtenidos se calcula la distancia de Hamming entre cada uno de los valores correspondientes a las imágenes modificadas y el hash de la imagen original. Esta métrica mide la cantidad de posiciones en las que dos cadenas binarias difieren y resulta adecuada para los objetivos del trabajo, ya que permite identificar variaciones entre los valores de hash generados sin considerar la ubicación específica de los bits modificados. En este contexto, el interés se centra en detectar la existencia de diferencias entre los resúmenes generados, independientemente del peso relativo de los bits que hayan cambiado.

Los datos producidos durante la experiencia se almacenan en formatos abiertos, preferentemente CSV, debido a la facilidad que ofrecen para su procesamiento mediante distintos lenguajes de programación y herramientas de análisis de datos. En particular, se registran los valores de hash obtenidos para cada imagen y sus variantes, junto con las distancias de Hamming calculadas respecto de la imagen original.

Con el fin de garantizar la reproducibilidad del estudio, el procedimiento experimental se implementa mediante scripts desarrollados en Python utilizando bibliotecas ampliamente difundidas como Matplotlib (licencia compatible con BSD) y Pandas (licencia BSD de tres cláusulas). Asimismo, el análisis de los datos también puede realizarse utilizando el lenguaje R (licencia GPL). Las herramientas desarrolladas en el marco de este trabajo se encuentran disponibles públicamente en el repositorio:

https://github.com/Kemmotar28/Document-Images-Tools.git

Las principales herramientas desarrolladas para la realización de la experiencia se resumen en la Tabla 1. Estos programas permiten automatizar tanto la generación de las imágenes utilizadas en el estudio como el cálculo y almacenamiento de los valores de hashing perceptual obtenidos durante las pruebas.

Tabla 1: Herramientas desarrolladas para la experiencia

Herramienta	Función
generadorDeImagenesDeDocumentos.py	Generador de imágenes
comparaHashesPerceptuales.py	Cálculo y almacenamiento de hashes

Las herramientas fueron desarrolladas en Python y forman parte del conjunto de scripts disponibles en el repositorio del proyecto, lo que facilita la replicación del procedimiento experimental y la extensión de los experimentos en futuros estudios.

Para facilitar la replicación del procedimiento en distintos entornos de investigación, la generación de imágenes se realiza con la asistencia de modelos de lenguaje de gran escala (LLMs), listados en la Tabla 2, priorizando configuraciones que puedan ejecutarse con requerimientos de hardware moderados. Este enfoque permite reducir el costo computacional del proceso y favorece la posibilidad de reproducir los experimentos en distintos contextos.

Tabla 2. LLMs utilizados en la experiencia

LLM	Licencia
Nous-Hermes-2-Mistral-7B-DPO.Q4_0	Apache 2.0
Phi-3-mini-4k-instruct.Q4_0	MIT
Meta-Llama-3-8B-Instruct.Q4_0	Llama 3
DeepSeek-R1-Distill-Llama-8B-Q4_0	MIT

5. Desarrollo de la experiencia

En este estudio se seleccionó un conjunto de funciones de hashing perceptual con el propósito de analizar su posible utilización como mecanismo para verificar la integridad del contenido textual presente en imágenes de documentos. Para ello se consideraron dos propiedades fundamentales que, según señalan Hadmi et al. [8], deberían caracterizar a un hash perceptual cuando se emplea con este objetivo.

En primer lugar, si dos imágenes resultan perceptualmente equivalentes, la probabilidad de que produzcan el mismo valor de hash debería ser cercana a uno. Esta condición se expresa formalmente en la ecuación (1). En segundo lugar, cuando dos imágenes presentan diferencias perceptuales, la probabilidad de que generen el mismo hash debería aproximarse a cero, tal como se indica en la ecuación (2).

$$x \approx y \Rightarrow P\big(H(x) = H(y)\big) \approx 1 \quad (1)$$

$$x \neq y \Rightarrow P\big(H(x) = H(y)\big) \approx 0 \quad (2)$$

El estudio se llevó a cabo sobre un conjunto de 1000 imágenes de documentos textuales sintéticos generadas automáticamente mediante software en un entorno experimental controlado.

Para cada imagen original se generaron distintas variantes con modificaciones controladas. En particular, se introdujo una alteración semántica mínima consistente en el cambio del valor de un número dentro del texto del documento. Además, se aplicaron transformaciones que no modifican el contenido semántico del documento pero que son frecuentes en el procesamiento de imágenes. Entre estas se incluyen la generación de una versión comprimida en formato JPEG con una calidad del 75 % y una versión con reducción del brillo al 60 %.

A continuación, se calcularon los valores de hash correspondientes a cada imagen y sus variantes utilizando la biblioteca *imageHash* en Python (distribuida bajo licencia BSD de dos cláusulas). Las funciones consideradas en el análisis fueron las siguientes:

- hash promedio (aHash)
- hash perceptual (pHash)
- hash diferencial (dHash)
- hash wavelet (wHash)

Posteriormente se calculó la distancia de Hamming entre el hash correspondiente a la imagen original y los hashes generados para cada una de sus variantes, utilizando siempre la misma función de hashing. Los valores obtenidos, tanto los hashes como las distancias de Hamming, se registraron en archivos en formato CSV para facilitar su posterior procesamiento y análisis.

Como indicador básico del comportamiento de las funciones analizadas frente a modificaciones de las imágenes, se contabilizó el número de casos en los que la distancia de Hamming resultó distinta de cero. Este criterio permite identificar situaciones en las que una modificación de la imagen provoca un cambio en el valor del hash generado.

La Figura 2 muestra la distribución de las distancias de Hamming distintas de cero para las distintas funciones de hashing perceptual analizadas. En el caso de las modificaciones semánticas (representadas en azul en la figura), las distancias observadas tienden a concentrarse en valores reducidos. Este comportamiento indica que los algoritmos presentan

cierto grado de invariancia frente a modificaciones que alteran mínimamente el contenido textual del documento.

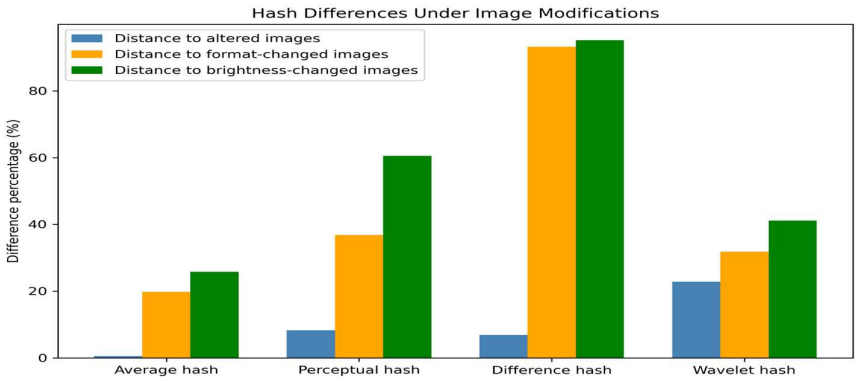

Figura 2. Distribución de los porcentajes de hashes diferentes

En relación con las transformaciones que no modifican el significado del texto, los resultados muestran que ninguno de los hashes evaluados satisface plenamente la propiedad de invariancia esperada. En el caso de la reducción del brillo, por ejemplo, aproximadamente la mitad de las imágenes produce valores de hash distintos para las funciones perceptual y wavelet, aun cuando el contenido textual permanece sin cambios. Para estas mismas condiciones, el hash diferencial genera diferencias en la gran mayoría de los casos (95,2 %), mientras que el hash promedio (aHash) exhibe un comportamiento relativamente más estable, manteniendo coincidencias con la imagen original en más del 70 % de las observaciones.

Una situación similar se observa al aplicar compresión JPEG sin alterar el contenido semántico del documento. En este caso, todas las funciones analizadas producen diferencias en los valores de hash respecto de la imagen original. La distribución de dichas diferencias resulta comparable a la observada en el experimento

En la tabla 3 presentamos el resumen del comportamiento de todas las funciones estudiadas.

Tabla 3: Detección de hashes diferentes entre las correspondientes a las imágenes modificadas y las originales.

Function / Comparison	Non zero distances	Percentages of non zero values
Original vs. Semantically Modified		
aHash	5	0,50%
pHash	83	8,30%
dHash	69	6,90%
wHash	228	22,80%

Original vs. JPEG Compression (75%)		
aHash	198	19,80%
pHash	368	36,80%
dHash	932	93,20%
wHash	318	31,80%
Original vs. Decreased Brightness (60%)		
aHash	258	25,80%
pHash	605	60,50%
dHash	952	95,20%
wHash	411	41,10%

Se observa que ninguno de los hashes evaluados cumple adecuadamente con la propiedad de invarianza frente a modificaciones no semánticas. En particular, en casi la mitad de las imágenes con brillo reducido, los hashes perceptual y wavelet producen valores distintos a los de la imagen original, aun cuando el contenido textual permanece inalterado. Para los mismos casos, el hash diferencial genera valores diferentes en la gran mayoría de las imágenes (95,2 %), mientras que el hash promedio presenta un comportamiento relativamente más estable, al coincidir en más del 70 % de los casos.

6. Discusión

El análisis realizado permite examinar con mayor detalle el comportamiento de distintas funciones de hashing perceptual cuando se aplican a imágenes que contienen documentos textuales. Los resultados obtenidos muestran que estos algoritmos presentan sensibilidades claramente diferenciadas frente a modificaciones en el contenido del texto, así como frente a transformaciones habituales en el procesamiento de imágenes, como cambios en el brillo o en el formato de almacenamiento.

En relación con la detección de modificaciones mínimas en el contenido textual, ninguna de las funciones analizadas logra generar valores de hash diferentes en la mayoría de los casos. Entre los algoritmos considerados, el hash wavelet muestra una mayor sensibilidad relativa, registrando diferencias en aproximadamente una cuarta parte de las imágenes que contienen modificaciones semánticas. No obstante, este comportamiento sigue siendo insuficiente para distinguir de manera confiable entre imágenes con contenido textual equivalente y aquellas que presentan alteraciones.

Por otra parte, los resultados evidencian que las funciones de hashing perceptual consideradas tampoco mantienen la invariancia esperada frente a transformaciones que no modifican el significado del texto. En particular, la reducción del brillo provoca discrepancias en los valores de hash en una proporción significativa de imágenes. En este escenario, dos de las funciones evaluadas (pHash y dHash) producen valores distintos en la mayoría de los casos,

mientras que las restantes (aHash y wHash) también registran diferencias en un número considerable de imágenes.

Estos resultados ponen de manifiesto una elevada sensibilidad de los algoritmos frente a transformaciones globales de luminancia, incluso cuando el contenido textual permanece intacto. En consecuencia, ninguna de las funciones analizadas resulta suficiente, de manera individual, para verificar la preservación de la integridad semántica de imágenes de documentos textuales.

Sin embargo, las diferencias observadas en el comportamiento relativo de los distintos algoritmos sugieren la conveniencia de profundizar el análisis en trabajos futuros. En particular, podría resultar de interés explorar estrategias que combinen múltiples funciones de hashing o examinar con mayor detalle las características específicas de los documentos en los que algunas de estas funciones ofrecen resultados más próximos a los esperados.

Más allá del desempeño particular de los algoritmos evaluados, la experiencia realizada muestra que el procedimiento experimental propuesto constituye una base metodológica adecuada para estudiar empíricamente el comportamiento de funciones de hashing perceptual aplicadas a imágenes de documentos. Durante el desarrollo del experimento se generó un volumen considerable de datos que no fue explotado completamente en el presente análisis, pero que abre la posibilidad de realizar estudios adicionales sobre el comportamiento de estos algoritmos frente a distintos conjuntos de imágenes.

7. Conclusiones

La experiencia desarrollada demuestra que es posible implementar estudios empíricos reproducibles orientados al análisis del comportamiento de funciones de hashing aplicadas a la protección de la integridad de imágenes que contienen documentos textuales. La utilización de datos sintéticos permite además sortear las dificultades asociadas a la obtención de grandes volúmenes de documentos provenientes de contextos reales, lo que facilita la construcción de entornos experimentales controlados y replicables.

En el caso analizado, los resultados obtenidos indican que las funciones de hash evaluadas no satisfacen plenamente los requisitos necesarios para ser consideradas mecanismos robustos de verificación de integridad en imágenes de documentos textuales. En particular, se observa que su comportamiento resulta limitado frente a modificaciones semánticas mínimas en el contenido textual y frente a transformaciones habituales en el procesamiento de imágenes que no alteran el significado del documento.

Durante el desarrollo del experimento se generó un conjunto de datos que no fue explotado completamente en el presente trabajo. Entre la información almacenada se encuentran los contenidos textuales originales utilizados para generar las imágenes sintéticas, lo que abre la posibilidad de realizar análisis adicionales. Por ejemplo, estos datos podrían emplearse para evaluar comparativamente el desempeño de distintas herramientas de reconocimiento óptico de caracteres (OCR) sobre el mismo conjunto de documentos.

Asimismo, los datos recolectados permiten realizar otros tipos de análisis que no fueron abordados en esta experiencia. Entre ellos se encuentran el estudio más detallado de la distribución empírica de los valores de hash generados por cada función o la exploración de representaciones gráficas de dichas funciones, cuyo análisis podría aportar indicios sobre su grado de uniformidad o sobre la probabilidad de aparición de colisiones.

En conjunto, el procedimiento experimental propuesto constituye una base metodológica que facilita la realización de evaluaciones empíricas reproducibles y comparables sobre técnicas de hashing perceptual aplicadas a imágenes de documentos textuales, contribuyendo a una comprensión más precisa de sus capacidades y limitaciones en escenarios de aplicación práctica.

Referencias

1. Wang, X., Yu, H.: How to break MD5 and other hash functions. In: Annual international conference on the theory and applications of cryptographic techniques. pp. 19–35. Springer (2005).
2. Tchórzewski, J., Jakóbik, A.: Theoretical and experimental analysis of cryptographic hash functions. J. Telecommun. Inf. Technol. 125–133 (2019).
3. Hasan, H.A., Al-Layla, H.F., Ibraheem, F.N.: A review of hash function types and their applications. Wasit J. Comput. Math. Sci. 1, 75–88 (2022).
4. Abramova, S., Böhme, R.: Detecting copy–move forgeries in scanned text documents. Electron. Imaging. 28, 1–9 (2016).
5. Pershin, I.: Evaluating Text Humanlikeness via Self-Similarity Exponent. In: ICLR 2025 Workshop on Building Trust in Language Models and Applications.
6. Swaminathan, A., Mao, Y., Wu, M.: Robust and secure image hashing. IEEE Trans. Inf. Forensics Secur. 1, 215–230 (2006).
7. Farid, H.: An overview of perceptual hashing. J. Online Trust Saf. 1, (2021).
8. Hadmi, A., Puech, W., Said, B.A.E.: Perceptual image hashing. Watermarking Vol. 2. 17 (2012).
9. Du, L., Ho, A.T., Cong, R.: Perceptual hashing for image authentication: A survey. Signal Process. Image Commun. 81, 115713 (2020).
10. Gangavarapu, T., Jaidhar, C.: An Empirical Study to Detect the Collision Rate in Similarity Hashing Algorithm Using MD5. In: 2019 International Conference on Data Science and Engineering (ICDSE). pp. 11–14. IEEE (2019).
11. Qu, C., Liu, C., Liu, Y., Chen, X., Peng, D., Guo, F., Jin, L.: Towards robust tampered text detection in document image: New dataset and new solution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5937–5946 (2023).
12. Owolabi, O.: Empirical studies of some hashing functions. Inf. Softw. Technol. 45, 109–112 (2003).
13. Luo, D., Liu, Y., Yang, R., Liu, X., Zeng, J., Zhou, Y., Bai, X.: Toward real text manipulation detection: New dataset and new solution. Pattern Recognit. 157, 110828 (2025).
14. Bausell, R.B.: The problem with science: The reproducibility crisis and what to do about it. Oxford University Press (2021).
15. Baker, M.: 1,500 scientists lift the lid on reproducibility. Nat. News. 533, 452 (2016).
16. Ioannidis, J.P.: Why most published research findings are false. PLoS Med. 2, e124 (2005).
17. Barba, L.A.: Defining the role of open source software in research reproducibility. Computer. 55, 40–48 (2022).
18. Ince, D.C., Hatton, L., Graham-Cumming, J.: The case for open computer programs. Nature. 482, 485–488 (2012). https://doi.org/10.1038/nature10836.
19. Stodden, V.: Enabling reproducible research: Open licensing for scientific innovation. Int. J. Commun. Law Policy Forthcom. (2009).
20. Fidler, F., Wilcox, J.: Reproducibility of scientific results. (2018).
21. Panda, S., Mishra, M.: Passive techniques of digital image forgery detection: developments and challenges. In: Advances in Electronics, Communication and Computing: ETAEERE-2016. pp. 281–290. Springer (2017).
22. Husain, A., Bakhtiari, M., Zainal, A.: Printed document integrity verification using barcode. J. Teknol. 70, (2014).
23. Tan, L., Hu, K., Zhou, X., Chen, R., Jiang, W.: Print-scan invariant text image watermarking for hardcopy document authentication. Multimed. Tools Appl. 78, 13189–13211 (2019).
24. Rhayma, H., Makhloufi, A., Hamam, H., Hamida, A.B.: Semi-fragile watermarking scheme based on perceptual hash function (PHF) for image tampering detection. Multimed. Tools Appl. 80, 26813–26832 (2021).