

# Document semantic hashing for hybrid security Sébastien Eskenazi, Petra Gomez-Krämer, Jean-Marc Ogier - Laboratoire L3i - La Rochelle, France

## **Context: document security**



## State of the art and remaining challeng

### **SIGNED project** [1]

- Cut the document image in squares of 64 by 64 pixels
- Apply a Haar Discrete Wavelet Transform on each square
- Create a fuzzy hash
- Other undisclosed pre- and post-processing steps.

Strengths	Weaknesses	
Probability of false alarm <0,001	Cannot detect the replacement of dots by commas with an error <0,001	L t
Probability of missed detection <0.001 for the replacement of digits	Cannot detect a manipulation smaller than 64x64 pixels at 600dpi (42x42 required)	
Collision probability <a></a>	Throughput >5s per page	
Compatible with current scanners and printers	Hash size >4kB (between 4,8 and 170 kB)	

#### References:

[1]A. Malvido Garcia: Secure Imprint Generated for Paper Documents (SIGNED). Technical Report December 2010, Bit Oceans (2013) [2] S. Eskenazi, P. Gomez-Krämer, J.-M. Ogier: When document security brings new challenges to document analysis. International Workshop on Computational Forensics (IWCF), (2014)

			Hash	computation p
	ure paper ument			Document image
	ure digital			Geometric correction
	ure hybrid ument			Segmentation
je	S		Table analysis	Image analysis
	MINISTERIO DE LA PRESIDENCIA MINISTERIO DE LA PRESIDENCIA Vacional de Interoperabilidad en el ámbito de la Administración Electrónica. La interoperabilidad es la capacidad de los sistemas de información y de los procedimientos a los que éstos dan soporte de compartir datos y posibilitar el intercambio de información y conocimiento entre ellos. Resulta necesaria para la cooperación el desarrollo, la integración y la prestación de servicios onjúntos por la Administración de servicios conjúntos por			Document robust reconstruction
	as Administraciones publicas, para la ejecución de las diversas políticas publicas, iara la realización de diferentes principios y derechos; para la transferencia de ecnología y la reutilización de aplicaciones en beneficio de una mejor eficiencia; <b>jara la cooperación entre diferentes aplicaciones que habiliten nuevos</b> <b>jervicios</b> ; todo ello facilitando el desarrollo de la administración electrónica y de la iociedad de la información. En el ámbito de las Administraciones públicas, la consagración del derecho de los iudadanos a comunicarse con ellas a través de medios electrónicos comporta una obligación correlativa de las mismas. Esta obligación tiene, como premisas, la promoción de las condiciones para que <b>la libertad y la igualdad sean reales y</b> <b>afectivas</b> , así como la remoción de los obstáculos que impidan o dificulten el apercicio pleno del principio de neutralidad tecnológica y de adaptabilidad al orogreso de las tecnologías de la información y las comunicaciones, garantizando con ello la independencia en la elección de las alternativas tecnológicas por los ciudadanos, así como la libertad de desarrollar e implantar los avances tecnológicos en un ámbito de libre mercado.			Document hash computing
L S E E E E E C C C C C C C C C C C C C C	a <b>Ley 11/2007</b> , de 23 de junité de acceso electrónico de los ciudadanos a los tervicios públicos, reconoce el protagonismo de la interoperabilidad y se refiere a illa como uno de los aspectos en los que es obligado que las previsiones tormativas sean comunes y debe ser, por tanto, abordado por la regulación del istado. La interoperabilidad se recoge dentro del principio de cooperación en artículos 4, 8 y tiene un protagonismo singular en el título cuarto dedicado a la Dooperación entre Administraciones para el impulso de la administración electrónica. En dicho título el aseguramiento de la interoperabilidad de los sistemas / aplicaciones del órgano de cooperación en esta materia el COMITÉ SECTORIAL DE ADMINISTRACIÓN ELECTRÓNICA. A CONTINUACIÓN, EL ARTÍCULO 31 SE REFIERE A LA APLICACIÓN POR PARTE DE LAS ADMINISTRACIONES PÚBLICAS DE LAS MEDIDAS INFORMÁTICAS, TECNOLÓGICAS Y ORGANIZATIVAS, Y DE SEGURIDAD, QUE GARANTICEN UN			Document signature
	IDECUADO NIVEL DE INTEROPERABILIDAD TÉCNICA, SEMÁNTICA Y IRGANIZATIVA Y EVITEN DISCRIMINACIÓN A LOS CIUDADANOS POR RAZÓN DE		954f7d	96502b5c5fe2e98a

Document secured by he SIGNED project, it takes 6 2D-barcodes to embed the signature

#### For further information



13i.univ-larochelle.fr

# **Conclusion: We need to study** and improve the robustness of document analysis algorithms.

### **Best case scenario results:**

- Accuracy : 99,83%
- Probability of false positive: 53% !!!
- Collision probability : 0,2%
- Other criteria are OK (throughput, digest size...)

### **False positives:**

- Occurs when two identical documents have different digests
- Related to the robustness of the OCR algorithm

$$P_{i} = \sum_{j=1}^{n_{i}} \left( \frac{n_{ij}}{\sum_{k=1}^{n_{i}} n_{ij}} \times \frac{n_{ij} - 1}{\sum_{k=1}^{n_{i}} n_{ij} - 1} \right)$$

## process



#### Text OCR processing



5045bca7f5

Computed as a random draw of two copies of the same document:



*n* documents *n<sub>i</sub>* digests per document - Each digest is present *n<sub>ii</sub>* times

# What is a semantic hash?

The second secon

Generation and validation of atlificial document image datas F sciences, N. Nayri, K. Romainyol, F. Gomez-Aramer, J. Charal Jummen, J-M. Oginer Jul - Universite de La Rochelle Pole Sciences et rechnologics, Avenue K. Crepeau, 17642 La Roch Bobatien, eskemanismiv-L.fr

00110101001 Semantic hash

# Text OCR processing [2]

**Test of Tesseract on 28512 document images** 

#### **Dataset:**

- 22 texts
- 6 fonts

\_

- 3 font sizes
- 4 font emphasis
- 3 printers
- 3 scanners
- 3 scanning resolutions



### **Test protocol:**

#### Character

Empty line

Tabulation and s

- (long hyphen)
- ` (left and righ

","," (left and ri double apostrop

I, I, 1 (capital i, 1 alphabet, numb

O (capital o)

fi (ligature)

fl (ligature)







#### - Run Tesseract on the document image - Post processing : alphabet reduction - Compute the SHA-256 hash of the document

Alphabet reduction				
	Replacement			
	Removed			
space	Removed			
	- (short hyphen)			
t apostrophes)	' (centered apostrophe)			
ght quotes, phe)	" (centered quote)			
2th letter of the er 1)	(vertical bar)			
	0 (zero)			
	fi (two letters f and i)			
	fl (two letters f and l)			