

الجمهورية الشعبية الديمقراطية الجزائرية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
المدرسة العليا للإعلام الآلي 8 ماي 1945 - سيدي بلعباس
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbès



Master's Thesis

To obtain the diploma of Master's Degree

Field of Study: **Computer Science**

Specialization: **Systèmes d'Information et Web (SIW)**

Theme

Document Forgery Detection & Authenticity Verification

Presented by
KADRI Ayyoub
BOUKRA BETTAYEB Abdellah

Defended on: **JULY, 2026**
In front of the jury composed of

Mr. BEDJAOUI Mohammed
Mr. MALKI Abdelhamid
Mr. LOUHAB Zouaoui

President of the Jury
Thesis Supervisor
Examiner

Abstract

Document forgery has become an increasingly important challenge in the digital era, where administrative, legal, financial, and identity documents are frequently scanned, exchanged, stored, and verified through digital platforms. The availability of advanced image editing tools, low-cost scanning and printing devices, and recent artificial intelligence-based generation techniques has made document manipulation easier, more realistic, and harder to detect. Forged documents may involve subtle modifications such as changing a name, date, amount, signature, stamp, or identity field, often affecting only a very small region of the document image.

This thesis presents a structured study of document forgery detection and authenticity verification techniques. It first introduces the fundamental concepts of document forensics, including common forgery types such as copy-move, splicing, and insertion, as well as the distinction between active and passive verification methods. It then reviews traditional detection approaches based on handcrafted features, including texture, noise, keypoints, frequency-domain artifacts, and edge information. The thesis further examines modern deep learning-based methods, with particular attention to convolutional neural networks, transformer-based architectures, document-specific detection models, and edge-aware approaches.

Special attention is given to recent benchmark-based findings showing that document forgery detection remains an unsolved problem in realistic conditions. In particular, current methods often suffer from poor cross-domain generalization and threshold calibration issues, especially when forged regions occupy only a small fraction of the image. The thesis also discusses emerging challenges such as receipt forgery, identity document manipulation, and AI-generated document forgeries. Through this analysis, the work highlights the need for robust, generalizable, and calibration-aware authenticity verification systems capable of detecting subtle manipulations across diverse document types.

Keywords— Document Authenticity, Forgery Detection, Large Language Models (LLMs), Information Extraction, OCR, Staged Framework, Siamese Neural Networks (SNNs)

Résumé

La falsification de documents est devenue un défi de plus en plus important à l'ère numérique, où les documents administratifs, juridiques, financiers et d'identité sont fréquemment numérisés, échangés, stockés et vérifiés au moyen de plateformes numériques. La disponibilité d'outils avancés de retouche d'images, de dispositifs de numérisation et d'impression à faible coût, ainsi que de techniques récentes de génération fondées sur l'intelligence artificielle, a rendu la manipulation de documents plus facile, plus réaliste et plus difficile à détecter. Les documents falsifiés peuvent comporter des modifications subtiles, telles que le changement d'un nom, d'une date, d'un montant, d'une signature, d'un cachet ou d'un champ d'identité, qui ne touchent souvent qu'une très petite région de l'image du document.

Ce mémoire présente une étude structurée des techniques de détection de falsification de documents et de vérification de leur authenticité. Il introduit d'abord les concepts fondamentaux de la criminalistique documentaire, notamment les types courants de falsification tels que le copier-déplacer, le photomontage et l'insertion, ainsi que la distinction entre les méthodes de vérification actives et passives. Il passe ensuite en revue les approches traditionnelles de détection fondées sur des caractéristiques conçues manuellement, notamment la texture, le bruit, les points clés, les artefacts du domaine fréquentiel et les informations de contour. Le mémoire examine également les méthodes modernes fondées sur l'apprentissage profond, avec une attention particulière portée aux réseaux de neurones convolutifs, aux architectures à transformeurs, aux modèles spécialisés pour les documents et aux approches sensibles aux contours.

Une attention particulière est accordée aux résultats récents issus de bancs d'essai, qui montrent que la détection de falsification de documents demeure un problème non résolu dans des conditions réalistes. En particulier, les méthodes actuelles souffrent souvent d'une faible généralisation entre domaines et de problèmes de calibration des seuils, surtout lorsque les régions falsifiées n'occupent qu'une faible fraction de l'image. Le mémoire aborde aussi des défis émergents tels que la falsification de reçus, la manipulation de documents d'identité et les falsifications générées par l'intelligence artificielle. À travers cette analyse, ce travail met en évidence la nécessité de systèmes de vérification d'authenticité robustes, généralisables et sensibles à la calibration, capables de détecter des manipulations subtiles dans divers types de documents.

Mots-clés— Authenticité des documents, détection de falsification, grands modèles de langage (LLMs), extraction d'information, OCR arabe, cadre par étapes, réseaux de neurones siamois (SNNs)

الملخص

أصبحت عملية تزوير الوثائق تحدياً متزايد الأهمية في العصر الرقمي، حيث تُسح الوثائق الإدارية والقانونية والمالية ووثائق الهوية وتُتبادل وتُخزن وتُتحقق عبر منصات رقمية. وقد جعل توفر أدوات متقدمة لتحرير الصور، وأجهزة مسح وطباعة منخفضة التكلفة، وتقنيات حديثة للتوليد المعتمد على الذكاء الاصطناعي، التلاعب بالوثائق أسهل وأكثر واقعية وأصعب اكتشافاً. وقد تتضمن الوثائق المزورة تعديلات دقيقة مثل تغيير اسم أو تاريخ أو مبلغ أو توقيع أو ختم أو حقل من حقول الهوية، وغالباً ما تؤثر هذه التعديلات في منطقة صغيرة جداً من صورة الوثيقة.

تقدم هذه الأطروحة دراسة منظمة لتقنيات كشف تزوير الوثائق والتحقق من أصالتها. وتبدأ بعرض المفاهيم الأساسية للجنايات الوثائقية، بما في ذلك الأنواع الشائعة للتزوير مثل النسخ والنقل، والتركيب، والإدراج، إضافة إلى التمييز بين طرق التحقق النشطة والسلبية. ثم تستعرض الأساليب التقليدية للكشف المعتمدة على خصائص مصممة يدوياً، مثل النسيج، والضجيج، والنقاط المفتاحية، وآثار المجال الترددي، ومعلومات الخواف. كما تتناول الأطروحة الأساليب الحديثة القائمة على التعلم العميق، مع تركيز خاص على الشبكات العصبية الالتفافية، ومعماريات المحولات، والنماذج المخصصة للوثائق، والطرائق الحساسة للخواف.

تولى عناية خاصة للنتائج الحديثة المبنية على مجموعات قياس مرجعية، والتي تُظهر أن كشف تزوير الوثائق لا يزال مشكلة غير محلولة في الظروف الواقعية. وبوجه خاص، تعاني الأساليب الحالية غالباً من ضعف التعميم عبر المجالات ومن مشكلات في معايرة العتبات، ولا سيما عندما تشغل المناطق المزورة جزءاً صغيراً فقط من الصورة. كما تناقش الأطروحة تحديات ناشئة مثل تزوير الإيصالات، والتلاعب بوثائق الهوية، وتزوير الوثائق المولدة بالذكاء الاصطناعي. ومن خلال هذا التحليل، يبرز العمل الحاجة إلى أنظمة تحقق من الأصالة تكون متينة وقابلة للتعميم وواعية بالمعايرة، وقادرة على كشف التلاعبات الدقيقة عبر أنواع متنوعة من الوثائق.

الكلمات المفتاحية--- أصالة الوثائق، كشف التزوير، النماذج اللغوية الكبيرة، استخراج المعلومات، التعرف الضوئي على الحروف العربية، إطار عمل مرحلي، الشبكات العصبية السيامية