Journal of Electronic Research and Application, 2026, Volume 10, Issue 1
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Application Research of Concept Bottleneck Model in Passport Printing Method Detection

Tianrui Qiu, Jiafeng Xu*

China People's Police University (Guangzhou), Guangzhou 510663, Guangdong, China

**Corresponding author: Jiafeng Xu, xujiafeng@cppu.edu.cn*

Abstract: With the increase in cross-border mobility, passports, as critical identity documents, require robust anti-counterfeiting security. While existing deep learning-based automatic detection methods achieve high accuracy, they lack interpretability. This paper introduces the Concept Bottleneck Model (CBM) to construct a transparent passport printing method detection framework. By defining interpretable intermediate concepts and integrating linear reasoning, the model significantly enhances reliability and debugging efficiency. The article systematically analyzes the advantages, challenges, and future directions of this approach.

Keywords: Passport anti-counterfeiting; Concept bottleneck model; Explainable AI; Printing method detection; Human-machine collaboration

Online publication: February 12, 2026

1. Introduction

Against the increasingly complex backdrop of globalization and security dynamics, passports, as the core credentials for national sovereignty and exit and entry administration, have their security directly linked to the stability of national security and social order. With the advancement of counterfeiting technologies, especially the growing number of cases involving the production of highly simulated passports using advanced digital printing and microprinting techniques, how to distinguish the authenticity of passports by detecting their printing methods (e.g., differentiating between optical printing, laser engraving and thermal printing) has become a key technical challenge in the fields of border security inspection and forensic authentication [1].

To address the above bottlenecks, this paper proposes an innovative approach of introducing Concept Bottleneck Models (CBMs) to construct a passport printing method detection framework that not only maintains high detection accuracy but also features interpretability and traceability. The main contributions of this paper are as follows: it systematically conducts a qualitative analysis of the adaptability of CBMs in the specific scenario of passport detection, deeply explores their application value in real security inspection environments, and analyzes the potential challenges in terms of model migration, concept annotation costs and cross-device universality.

2. Literature review of related work

2.1. Overview of passport printing processes and anti-counterfeiting features

As the highest-level identity documents issued by countries, passports adopt far more complex printing processes than ordinary paper documents. The current mainstream passport printing processes mainly include laser printing, inkjet printing and intaglio printing, each with unique physical characteristics and anti-counterfeiting mechanisms. Laser printing transfers toner to paper surfaces using photothermal technology, typically producing features such as high-contrast images and a specific fine granular texture; inkjet printing forms patterns through the deposition of tiny ink droplets, characterized by natural color gradation and weak granularity; intaglio printing, by contrast, imprints ink onto paper via recessed relief plates, creating patterns with distinct three-dimensional tactile feel and glossiness. A thorough understanding of the core differences between these processes forms the foundation for constructing an effective detection model.

2.2. Brief analysis of existing detection technologies

Current passport anti-counterfeiting detection technologies are mainly divided into traditional physical feature-based methods and deep learning-based automatic detection methods. Traditional methods rely on expert experience and specialized equipment to judge authenticity by analyzing the spectral reflectance of paper, watermark structures or intaglio tactile feel, yet they are limited by human factors and difficult to realize large-scale automation. While deep learning models perform excellently in the field of image recognition and have significantly improved detection speed and accuracy, their excessive abstraction capability for intermediate features results in a lack of interpretability, making it difficult to meet the stringent requirements for "trustworthiness" in safety-critical fields such as passport anti-counterfeiting [2].

2.3. Necessity of explainable artificial intelligence in safety-critical fields

In safety-critical fields such as passport detection, the application of technology is not only about pursuing high accuracy but, more crucially, ensuring the reliability and traceability of results. Traditional deep learning models often function as "black boxes": even if a model yields an accurate result, the lack of explanation means that when the model makes a misjudgment or is attacked by adversarial examples, it is impossible to quickly locate the root cause of the problem. In passport authentication, examiners need to understand the specific basis for the model's determination of "inkjet printing" or "laser printing" to conduct secondary verification. Therefore, the introduction of Explainable Artificial Intelligence (XAI) is an urgent priority, as it can provide the logical chain of model decision-making, thereby enhancing examiners' trust in the system and improving the security of the overall anti-counterfeiting system.

2.4. Core ideas and development context of concept bottleneck models: From concept learning to explainable reasoning

Concept Bottleneck Models (CBMs) are a type of deep learning architecture that integrates both interpretability and predictive ability, whose core idea is to design the model's hidden layers as a set of explicit human-interpretable "concepts" [3]. Specifically, a CBM first maps an input image to a set of predefined concept vectors (e.g., "paper glossiness", "ink granularity"), which directly correspond to the key features of passport printing processes. The model then makes the final prediction based on these concept vectors. This structure achieves a leap from "pixel-level features" to "concept-level features", rendering the model's intermediate decision-making process fully transparent. The proposal of CBMs marks the evolution of artificial intelligence research from "pure concept learning" to "concept-based explainable reasoning", providing a theoretical foundation for solving the interpretability bottleneck of deep learning models in passport detection.

3. Construction of a framework for applying concept bottleneck models to passport printing detection

3.1. Qualitative description of the overall architecture

The detection framework proposed in this section technically follows the standard workflow of Concept Bottleneck Models (CBMs) to realize transparent judgment of passport printing methods. Specifically, the system first receives high-resolution scanned or captured images of passports (input images), which contain the original physical information related to anti-counterfeiting features. The images then enter the concept extraction layer, where the model does not directly output category labels but extracts a set of intermediate features corresponding to human anti-counterfeiting experience, these features are referred to as "concepts". These concepts are subsequently aggregated and regularized through the concept bottleneck layer to form a set of structured concept vectors. Finally, the task prediction layer performs simple logical reasoning based on these concept vectors and outputs the final classification result of the printing method. Through the "bottleneck", this architecture maps complex image features to interpretable concepts, achieving a qualitative leap from a "black box" to a "transparent box".

3.2. Definition and design of core concepts

In passport anti-counterfeiting, concept design is crucial and must balance machine recognition with human examiners' experience. Targeting different printing processes, we define four core concepts:

(1) "Ink diffusion morphology" (capturing inkjet's natural diffusion edges, contrasting laser printing's sharp toner-fused edges);

(2) "Dot matrix structure" (a key feature distinguishing inkjet spray dots from laser powder deposition dots);

(3) "Gloss texture pattern" (reflecting the metallic luster of intaglio-printed paper vs. ordinary paper's matte texture under light);

(4) "Character edge sharpness" (quantifying character contour clarity and attenuation across printing methods). Derived from anti-counterfeiting experts' long-term empirical observations, these concepts ensure professional depth in the model's interpretable outputs.

3.3. Qualitative approach to concept extraction

To extract the above-defined concepts efficiently and accurately, the model adopts a multi-source fusion technical approach in the concept extraction layer. Specifically, a deep convolutional neural network (e.g., ResNet) pre-trained on large-scale image datasets is first used to extract general visual features, a step that ensures the model has strong feature perception capability [4]. Subsequently, in response to the specific needs of passport anti-counterfeiting, small domain knowledge-driven network branches are designed, which are fine-tuned specifically for microscopic features such as "glossiness" or "ink diffusion" to capture delicate physical differences. In addition, we introduce traditional feature engineering methods, such as Gabor filters or gray-level co-occurrence matrices, to assist in extracting “dot matrix structure” or “texture pattern”. Through the synergy of these approaches, the system can map complex image pixels to high-dimensional concept attribute vectors, laying a solid data foundation for subsequent reasoning.

3.4. Qualitative analysis of concept-task reasoning

Following feature abstraction in the concept bottleneck layer, the system proceeds to decision-making. The task prediction layer employs simple interpretable algorithms like linear or logistic regression, analyzing the weight contribution of each dimension in the concept vector: for instance, high “ink diffusion morphology” and low “gloss texture pattern” scores indicate inkjet printing, while extremely high “character edge sharpness” paired with prominent “gloss texture pattern” suggests laser or intaglio printing. The reasoning model’s simplicity enables direct access to each concept’s contribution to the final decision, ensuring transparent decision-making. This enhances model trustworthiness, allows examiners to conduct secondary verification via concept explanations, and significantly reduces misjudgment risks.

4. Qualitative analysis of application advantages and value

4.1. Improved interpretability and trustworthiness

The greatest advantage of CBMs lies in their intrinsic “interpretable” attribute. Although traditional deep learning models achieve high accuracy, their decision-making process is a completely opaque black box for humans, which is unacceptable in the authentication of safety-critical documents such as passports [5]. In contrast, CBMs map complex pixel-level features directly to specific physical features by imposing constraints that the model’s hidden layers must output concepts consistent with human cognition. This mapping process can be directly observed and reviewed by domain experts, making each judgment of the model based on verifiable evidence, greatly improving the trustworthiness of the entire detection system and avoiding legal and security risks caused by the “black box” nature.

4.2. Human-machine collaboration and knowledge fusion

CBMs are naturally designed to support “human-machine collaboration”. Since the model outputs human-interpretable concepts, this provides an intuitive intervention point for domain experts. If an expert finds that the model identifies certain specific concepts inaccurately, they can directly intervene and correct the concept without retraining the entire deep network. This interactive iterative optimization process not only can rapidly improve the performance of the model but also promotes the in-depth integration of machine learning algorithms with traditional anti-counterfeiting expertise, enabling the system to evolve continuously with the accumulation of expert experience.

4.3. Potential in data efficiency and generalization ability

By learning the intermediate concept layer, CBMs can alleviate the reliance on large-scale labeled data to a certain extent. Unlike traditional end-to-end learning that requires a large number of labeled samples to capture complex high-dimensional features, CBMs can use the definition of concepts by domain experts to guide the learning process [6]. This means that even when facing new counterfeiting methods or variants, the model can identify them well as long as these variants exhibit obvious anomalies in the concept space, thus demonstrating good generalization ability. This has important practical value for scenarios such as passport anti-counterfeiting detection that need to respond to constantly upgraded counterfeiting technologies.

4.4. Debugging and error diagnosis

CBMs have significant advantages in troubleshooting. When a traditional deep network malfunctions, engineers often struggle to locate the root cause of the error due to the abstraction and high dimensionality of internal features, resulting in an extremely complex debugging process. On the contrary, the decision-making of CBMs is based on a series of explicit concepts. If the system makes an identification error, engineers can directly check which concept (e.g., dot matrix structure or ink diffusion morphology) is misidentified, leading to the final misjudgment. This concept-based error localization mechanism not only accelerates the maintenance and iteration of the system but also significantly reduces the technical costs of long-term operation.

5. Challenges and future outlook

5.1. Existing challenges

Although Concept Bottleneck Models (CBMs) enhance deep learning interpretability via a human-interpretable concept layer, their application for passport anti-counterfeiting faces two key challenges. First, defining the concept system is highly complex: key anti-counterfeiting features of passport backgrounds correlate with continuous physical variables (e.g., glossiness, ink diffusion morphology) that vary significantly with illumination. This creates a dilemma: overly broad concepts reduce discriminability, while overly strict definitions drive up annotation costs, cause concept overlap and interpretation conflicts, making it hard to develop a concept set that covers key printing process features while enabling stable separation in images. Second, concept extraction reliability directly determines interpretability validity. Real-world passports often have non-ideal factors (wear, stains, uneven illumination); misjudgments by concept extractors in complex backgrounds trigger cascading errors that distort or collapse interpretive results, necessitating improved robustness and consistency of extractors against noise, distortion and background interference.

5.2. Future outlook

Future passport detection systems will evolve from closed black boxes into interactive, continuously learning platforms. First, they will adopt adaptive concept learning—automatically expanding or adjusting concept libraries for emerging counterfeiting methods, replacing rigid manual definitions. Second, CBMs will integrate with other XAI technologies (e.g., real-time visualization of “glossiness” response regions in images, adversarial example analysis to enhance model security). Ultimately, this technology will expand beyond passport detection to anti-counterfeiting for ID cards, driver’s licenses, certificates and bills, building a universal, interpretable AI ecosystem for document authentication.

6. Conclusion

The introduction of CBMs drives the transformation of passport printing method detection from “black box” judgment to “white box” reasoning. Compared with traditional deep learning, CBMs map pixel features to interpretable physical attributes (e.g., ink diffusion, gloss texture) through the concept layer, making the judgment basis traceable and verifiable, improving system credibility, and providing support for audit, compliance and error correction. However, its implementation cannot rely solely on algorithm optimization; it still requires in-depth collaboration between computer vision and document anti-counterfeiting experts to jointly define key concepts and conduct continuous calibration in iteration. Only in this way can a document authentication system with both intelligence and security be constructed.

Funding

2025–2026 China People's Police University Student Science and Technology Innovation Program Project

Disclosure statement

The authors declare no conflict of interest.

References

[1] Weeraratna T, 2024, Beyond Borders: The Art and Science of Detecting Travel Document Forgeries. International Journal of Forensic Sciences, 9(4): 1–4.

[2] Mohit M, 2016, The Evolution of Deep Learning: A Performance Analysis of CNNs in Image Recognition. International Journal of Advance Research in Education and Technology, 3(6): 2029–2038.

[3] Stropeni A, Enhancing Interpretability in Visual Anomaly Detection Through Concept Bottleneck Models, thesis, University of Padua.

[4] Dhillon A, Verma G, 2020, Convolutional Neural Network: A Review of Models, Methodologies and Applications to Object Detection. Progress in Artificial Intelligence, 9(2): 85–112.

[5] Shafik W, 2026, The "Black Box" Problem: Lack of Transparency in AI Decision-Making. Springer: 167–186.

[6] Srivastava D, Yan G, Weng L, et al., 2024, VLG-CBM: Training Concept Bottleneck Models with Vision–Language Guidance. Advances in Neural Information Processing Systems, 37: 79057–79094.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.