Obfuscation of Sensitive Text in Audiovisual Content Using AI

Kexin Jiang-Chen (✉)[0009−0005−2492−6531] and Cèsar Ferri[0000−0002−8975−1120]

VRAIN, Universitat Politècnica de València, València, Spain
kjiache@etsinf.upv.es, cferri@dsic.upv.es

Abstract. The digital revolution has led to an increase in audiovisual content across platforms, creating new challenges for privacy protection. Sensitive information, such as personal identifiers, financial data or contact information, frequently appears in images and videos, often unintentionally. These accidental disclosures can lead to serious privacy breaches or misuse of personal data. To address this issue, we present an automated solution for detecting and obscuring sensitive text in multimedia content, with particular focus on Spanish-language educational materials. Our system combines Microsoft Presidio's advanced Natural Language Processing (NLP) capabilities for Personally Identifiable Information (PII) detection with Tesseract Optical Character Recognition (OCR) text extraction from visual media. Detected sensitive content is then obfuscated using advanced image processing techniques, ensuring privacy protection while maintaining the visual quality of the multimedia. This integrated approach provides an effective, efficient method for protecting personal data in multimedia applications without compromising usability.

Keywords: Computer Vision · Obfuscation · Personally Identifiable Information.

1 Introduction

The rapid growth of audiovisual content on digital platforms has transformed information sharing while creating new privacy vulnerabilities. Personal data (names, addresses, phone numbers) frequently appears unintentionally in educational materials, social media, and streaming content, risking exposure under regulations like GDPR [1]. This is especially relevant in contexts such as tutorials and instructional videos created by educators, which may contain sensitive information, requiring careful handling to prevent privacy breaches. Current solutions face some limitations as manual review is inefficient, pure machine learning requires excessive resources, and rule-based systems lack flexibility.

We present a hybrid system combining Microsoft Presidio [7] for PII detection (using both rule-based and Machine Learning (ML) approaches) with Tesseract OCR [4] for text extraction from multimedia. The framework automatically obscures sensitive Spanish-language text via image processing while

preserving content quality. The system is customizable based on user-defined criteria and open-source, available on GitHub[1], and a demonstration video at link[2].

2 Related Work

The identification and obfuscation of sensitive information in multimedia is a growing research area [2] [6] and approaches generally fall into two categories: machine learning-based and rule-based methods.

Rule-based methods, such as spaCy [3] and Stanford NLP [10] offer transparency but struggle with multimedia variability. Microsoft Presidio bridges this gap through customisable rule patterns, though it remains limited for novel data types. Machine learning models, including CNNs and RNNs, excel in generalising across diverse data types, but can be inaccurate in ambiguous cases and require large datasets and computational resources [9].

For text extraction, Tesseract OCR [4] remains the open-source standard despite challenges with low-quality inputs, while commercial APIs (e.g., Google Vision) offer improved accuracy at higher costs. Complementary work in visual anonymisation (e.g., FaceNet [8]) focuses on anonymising visual content by blurring facial features, but does not address text-based obfuscation. Our system uniquely integrates Presidio's hybrid detection with Tesseract's extraction capabilities, creating a unified solution for multimedia privacy that handles both textual and contextual challenges.

3 Methodology

The proposed methodology is primarily designed to identify and obfuscate sensitive textual content in videos, but it also works effectively for static images.

Video frame extraction. The video is first broken down into individual frames at a predetermined frame rate. These frames serve as input for further analysis. Frame extraction ensures that each moment of the video is thoroughly scanned for sensitive content, allowing for a comprehensive detection and obfuscation process.

Text detection and preprocessing. We employ the OpenCV library [5], which provides a range of image processing functions to enhance text detection, including grayscale conversion, noise reduction, thresholding, and morphological operations (dilation/erosion) to improve OCR accuracy under challenging conditions such as motion blur or low contrast.

Text recognition. The preprocessed frames undergo OCR analysis using Tesseract, which detects character patterns and converts them to machine-readable text while recording precise spatial coordinates. This dual output, both

[1] https://github.com/Kexinjc/Text-obfuscation-in-videos
[2] https://drive.google.com/file/d/1ThsYGQc7qCIm7IizW4JTeGm7Jes5Wflt

textual content and positional data (bounding boxes), serves two key functions: (1) enabling NLP-based sensitivity analysis through Presidio, and (2) guiding targeted obfuscation by identifying exact regions requiring blurring. The bounding boxes maintain visual context during redaction, ensuring only sensitive elements are modified while preserving surrounding content integrity.

Sensitive information identification. We used Microsoft Presidio, a service designed to identify PII in text, employing a combination of predefined and custom PII recognisers primarily based on machine learning models. These recognisers detect sensitive information using techniques like Named Entity Recognition (NER) and heuristic methods. Presidio is capable of identifying a wide range of PII, such as names and locations, in multiple languages, including Spanish and English, which are key to this project.

In addition to the built-in recognisers, we extended Presidio with custom rule-based recognisers to detect sensitive information specific to our needs, such as dates of birth, Spanish ID numbers (DNI), phone numbers, addresses, and Spanish postal codes using regex patterns.

Additionally, we integrated contextual rules within the recognisers to improve detection accuracy by analysing the surrounding text to determine the likelihood of a word or phrase being sensitive. For instance, if a potential PII entity appears next to key phrases like "Date of Birth:", "Address:", or "Phone:", its probability of being classified as sensitive increases. This contextual approach reduces false positives and enhances the reliability of the detection process.

Obfuscation. Once sensitive content is identified, the tool applies obfuscation techniques. This process involves drawing bounding boxes around the sensitive text and applying a Gaussian blur filter to that area, effectively masking the text while preserving the overall integrity of the video frame. This approach ensures that the obfuscated content remains natural-looking, minimising distractions for viewers.

Frame difference calculation. Processing every frame in a video independently can be computationally expensive. To optimise performance, we implemented a frame differencing technique that calculates the percentage of change between consecutive frames. If the change is below a predefined threshold, the system reuses the bounding boxes from the previous frame, skipping redundant OCR and PII detection steps. This significantly reduces processing time, particularly in videos with minimal motion or static backgrounds, without compromising detection accuracy. The default threshold is 2%, chosen as a good trade-off between speed and precision, but it can be adjusted by the user as needed.

Post-processing and video reconstruction. After the obfuscation is applied to all frames, the modified frames are recompiled into a video format. The post-processing step ensures the integrity and quality of the original video are preserved while obfuscating the sensitive content.

4 Application

To make the system user-friendly and accessible, a Command Line Interface (CLI) is provided for managing the anonymisation process. This CLI supports both images and videos, using separate pipelines optimised for each input type. Users can customise the anonymisation process using various flags for flexible detection and obfuscation of sensitive content. The *-i* flag specifies the input video or image, and *-o* sets an optional output name. Use *-r* to select specific recognisers or *-e* to exclude them. The *-f* flag obfuscates specific words, while *-u* unmasks information to remain visible. The *-v* flag enables verbose mode for detailed process output, and *-t* sets the threshold for reusing bounding boxes based on frame-to-frame changes.

Figure 1 shows an example of a university intranet image with sensitive information obfuscated after being processed through the command line.

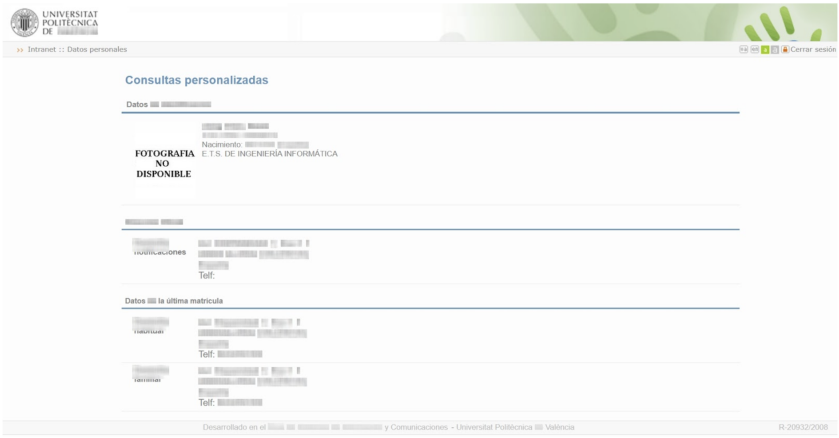

Fig. 1. University intranet image with sensitive information obfuscated after being processed through the command line interface.

5 Conclusions and Future Work

This paper presented a video processing system that detects and obfuscates sensitive text by combining Tesseract OCR and Microsoft Presidio. The system employs a frame differencing technique to optimise efficiency, reducing redundant processing and enhancing its practicality for non-real-time applications. It effectively preserves video integrity while obfuscating sensitive content, making it suitable for privacy and data protection use cases. The system's flexibility allows for extensive customisation, adapting to different content types and specific privacy requirements.

Future development will focus on three key areas:

1. **Detection improvements**: Integrating transformer-based OCR models for better handling of distorted text and extending recognisers to cover financial data, handwritten notes, and graphical PII (logos, signatures).
2. **Performance optimisation**: Implementing GPU acceleration and parallel processing to enable near real-time operation.
3. **Usability enhancements**: Adding multilingual support (starting with EU languages), cloud deployment options, and an interactive web interface to complement the existing CLI.

6 Acknowledgments

We acknowledge support from: Cátedra de Inteligencia Artificial aplicada a la Administración Pública and grant CIPROM/2022/6 (FASSLOW), both funded by Generalitat Valenciana; and Spanish grant PID2021-122830OB-C42 (SFERA) funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe".

References

1. de Carvalho, R., Del Prete, C., Martin, Y.e.a.: Protecting Citizens' Personal Data and Privacy: Joint Effort from GDPR EU Cluster Research Projects. SN COMPUT. SCI. **1**(217) (2020)
2. Di Cerbo, F., Trabelsi, S.: Towards personal data identification and anonymization using machine learning techniques. In: New Trends in Databases and Information Systems. ADBIS 2018. Communications in Computer and Information Science, vol. 909, pp. 409–421. Springer, Cham (2018)
3. Explosion: spaCy: Industrial-strength NLP. https://spacy.io (2014), accessed: October 1, 2024
4. Google: Tesseract OCR. https://github.com/tesseract-ocr/tesseract (2014), accessed: October 1, 2024
5. Itseez: Open source computer vision library. https://github.com/itseez/opencv (2015)
6. Marulli, F., Verde, L., Marrone, S., Barone, R., De Biase, M.: Evaluating efficiency and effectiveness of federated learning approaches in knowledge extraction tasks. In: Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–6. IEEE (2021)
7. Microsoft: Presidio - data protection and de-identification SDK. https://github.com/microsoft/presidio (2018), accessed: October 1, 2024
8. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2015). https://doi.org/10.1109/cvpr.2015.7298682, http://dx.doi.org/10.1109/CVPR.2015.7298682
9. Sharif, M., Khan, M.A., Usman, M., Lali, M.I., Saba, T., Rehman, A., Micor, J.R., Khurram, S., Gadekallu, T.R.: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data **8**(1), 1–74 (2021). https://doi.org/10.1186/s40537-021-00444-8
10. Stanford: Stanford CoreNLP: A suite of core NLP tools. https://stanfordnlp.github.io/CoreNLP/ (2013), accessed: October 1, 2024