DataGuardian: AI-Powered Multi-Modal Visual and Textual Data Anonymization System

Karamjeet Singh Gulati, Nikita B. Emberi, Jason Yoo

ECS 235A: Computer Information security Project presentation

Fall 2024

Objective and Goals

In today's digital age, there is an increasing need to protect personally identifiable information (PII) present in multi-modal data, such as images and text, from exposure. The central question our project aims to address is:

“How can an AI-powered system effectively anonymize both visual and textual data simultaneously, while preserving the utility of the data for research and analysis purposes?”

Goals:

- **Privacy Protection**: Real-time anonymization of visual and textual PII with advanced AI models.
- **Innovation**: Unified system, 700ms response time, modular and adaptable design.
- **Applications**: Healthcare, social media, document security, and research.
- **Vision:** Balancing privacy and utility with scalable, user-friendly solutions.

Methodology: Data Preprocessing

1. Visual Data:

- Real-time frame capture via Gradio API (30 FPS).
- RGB-to-BGR transformation, resolution standardization, and lossless JPEG compression.
- Secure UUID-based storage with automated cleanup and error handling.

2. Text Data:

- Multilingual NLP support for six languages and a universal fallback model.
- Accurate and efficient text anonymization.

3. Integration:

- Combined visual and textual preprocessing for seamless anonymization.

Methodology: Architecture

1. **Three-Stage Pipeline:** Streamlined design for real-time data processing and advanced anonymization.

2. **Input Layer:**
   - Gradio-based interface for webcam streams and text prompts.
   - Preprocesses multi-modal inputs while ensuring data integrity.

3. **Core Processing:** Real-time AI-driven analysis with strict privacy controls.

4. **Output Layer:**
   - Managed chat interface presenting anonymized data.
   - Maintains privacy compliance and conversational context.

5. **Key Features:**
   - Modular design for easy maintenance and future extensions.
   - Robust foundation for handling sensitive visual and textual data.

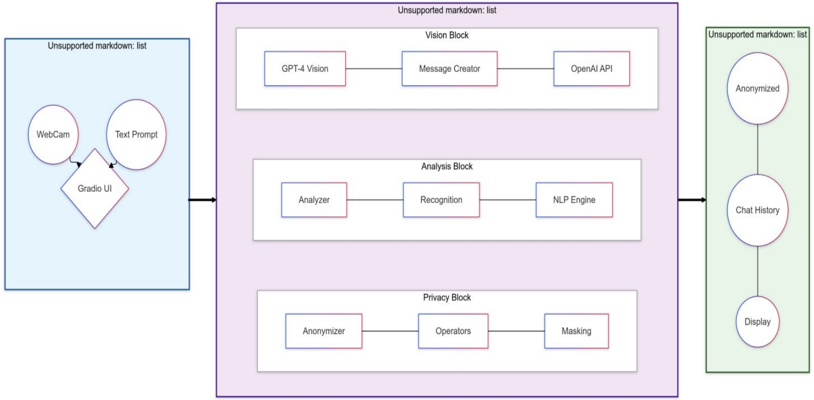


Figure: Three-stage pipeline integrating GPT-4 Vision and Presidio for input, processing, and output anonymization.

Methodology: Output Generation

Image Processing:

- We begin by processing visual data using Python modules that convert color spaces and correct image orientations. Each processed file is securely stored with error-resilient mechanisms.
- To enhance visual analysis, we integrate GPT-4 Vision with configurations like a token limit of 500 and a low temperature of 0.1, ensuring precise and focused outputs. Timeout management and fallback strategies handle potential errors.

Entity Recognition:

- For text data, using a confidence threshold of 0.2 and contextual enhancements, we ensure high accuracy in detecting entities such as emails, phone numbers, and personal names
- Identified entities are anonymized by replacing them with placeholders like -MASKED EMAIL RELATED-, preserving data privacy.

Output Generation:

- Processed data is formatted into JSON or HTML for structured output.
- We use an interactive interface that provides real-time feedback, displays chat functionality, and tracks chat history and image references. Temporary files are cleaned after processing to maintain data integrity and system efficiency.

Current Progress

Interface Development:

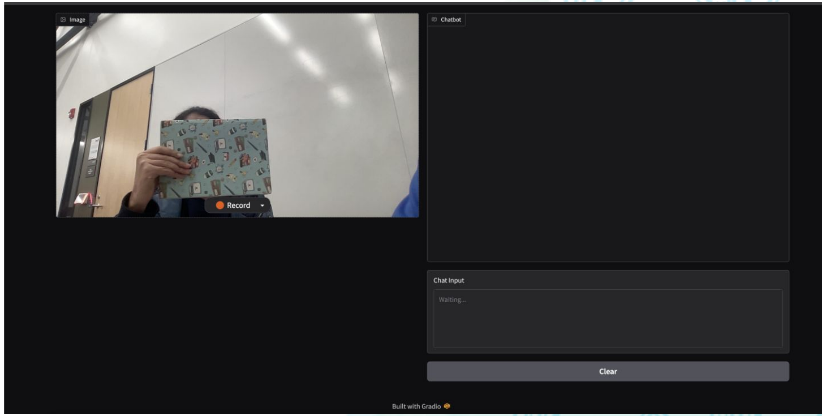
- Real-time webcam integration with Gradio.
- Responsive chat interface with multi-line input and history display.

Backend Enhancements:

- Integrated GPT-4 Vision for visual data analysis with response handling.
- Image processing pipeline includes color conversion, orientation correction, and secure UUID-based storage.

Privacy Mechanisms:

- Entity recognition identifies and anonymizes sensitive data (e.g., emails, phone numbers).
- Anonymization rules replace entities with placeholders like -MASKED PERSON RELATED-.



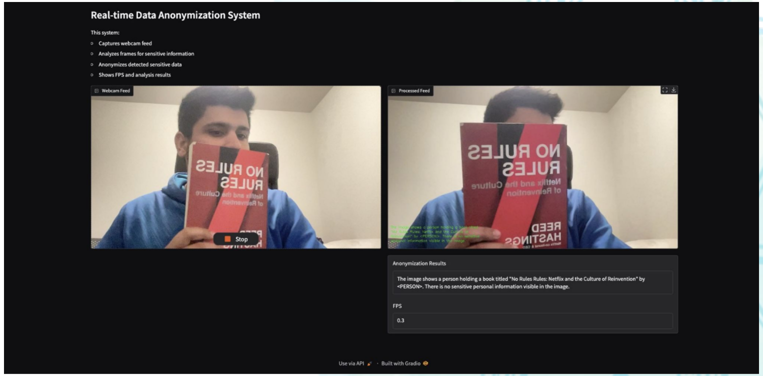
Current Progress

Multi-Language Support:

- Implemented models for six languages with optimized system parameters for processing.

Testing and Stability:

- Validated webcam capture, image processing, and file management.
- Server supports up to 10 threads for stable, scalable operations.



Real-time Data Anonymization System

This system:

- Captures webcam feed
- Analyzes frames for sensitive information
- Anonymizes detected sensitive data
- Shows FPS and analysis results

Anonymization Results

The image shows a person holding a book titled "No Rules Rules: Netflix and the Culture of Reinvention" by <PERSON>. There is no sensitive personal information visible in the image.

FPS

0.3

Current Results

1. Testing Setup:

- Tested on a local server (127.0.0.1:8800) with 10 threads and real-time processing.
- Evaluated response time (700ms), memory usage, entity recognition accuracy, and multi-language support.

2. Interface:

- Dual-panel Gradio interface: Webcam input on the left, interactive chat on the right.
- Real-time video streaming and frame capture using OpenCV.

3. Anonymization:

- Presidio framework detects sensitive data (e.g., names, emails) and replaces it with placeholders like `-MASKED PERSON RELATED-`.
- Errors in implementation currently limit full functionality.

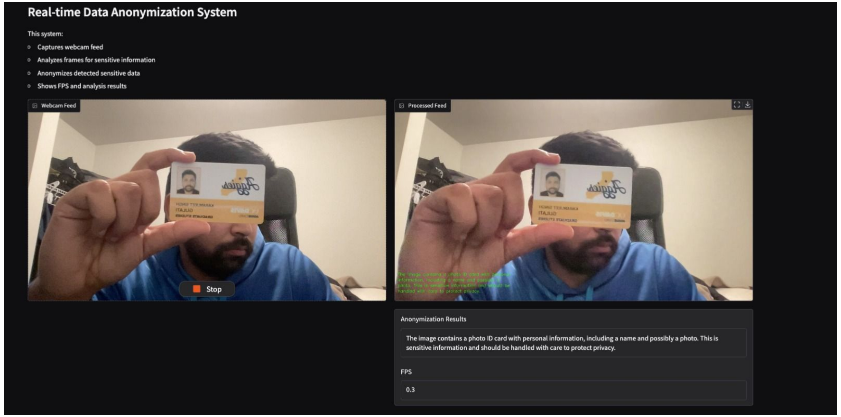
4. Modular Architecture:

- Python-based components using OpenCV, Presidio, and GPT-4 Vision ensure seamless integration.
- Designed for real-time privacy-focused processing with ongoing optimizations.

Current Results

1. Testing Outcomes:

- Validated: Frame capture, image quality, color conversion, and data storage.
- Preliminary: Anonymization patterns under review, with ongoing error resolution and refinements.


Webcam Feed
Stop
Processed Feed
The image contains a photo ID card with personal information, including a name and possibly a photo. This is sensitive information and should be handled with care to protect privacy.

Real-time Data Anonymization System

This system:

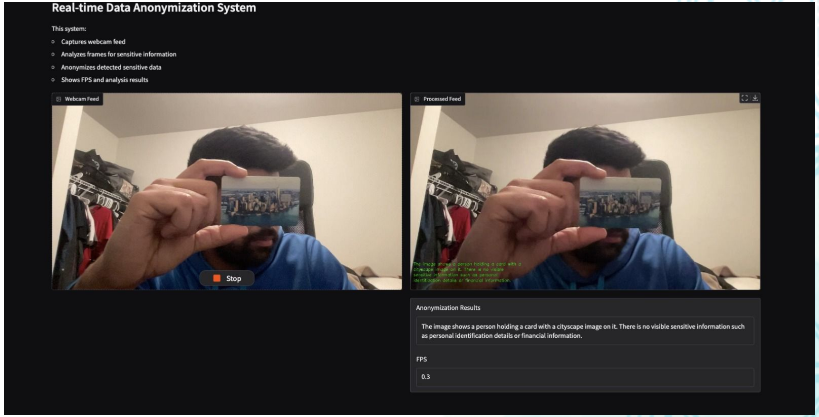
- Captures webcam feed
- Analyzes frames for sensitive information
- Anonymizes detected sensitive data
- Shows FPS and analysis results

Anonymization Results

The image contains a photo ID card with personal information, including a name and possibly a photo. This is sensitive information and should be handled with care to protect privacy.

FPS

0.3


Webcam Feed
Stop
Processed Feed
The image shows a person holding a card with a cityscape image on it. There is no visible sensitive information such as personal identification details or financial information.

Real-time Data Anonymization System

This system:

- Captures webcam feed
- Analyzes frames for sensitive information
- Anonymizes detected sensitive data
- Shows FPS and analysis results

Anonymization Results

The image shows a person holding a card with a cityscape image on it. There is no visible sensitive information such as personal identification details or financial information.

FPS

0.3

Next Steps

1. Development Completion:

- Add Gaussian blur for improved image anonymization.
- Finalize context-aware entity recognition and extend recognition patterns.

2. Performance Optimization:

- Enhance image compression, memory management, and thread pooling.
- Address latency issues, memory leaks, and improve error recovery.

3. System Expansion:

- Integrate additional language models and custom entity definitions.
- Implement advanced privacy rules and secure system access with user authentication.

4. Interface Enhancements:

- Add real-time statistics visualization and configuration options.
- Enable batch processing and improve the presentation of anonymized data.

Next Steps

5. Long-Term Vision:

- Use machine learning for automated rule generation and privacy policy compliance (GDPR/HIPAA).
- Deploy the system to the cloud for scalability and explore advanced anonymization techniques.