

CASE STUDY

Anti-Spoofing Video Datasets: Enhancing AI Security Against Fraudulent Attacks

Leveraging Shaip's expertise in high-quality dataset curation to support anti-spoofing AI models with diverse, real-world video data.



COMPANY

A leading AI security and biometrics company specializing in fraud prevention and anti-spoofing solutions.



INDUSTRY

AI Security, Biometrics, Fraud Prevention



USE CASE

Anti-Spoofing AI Model Training with Real and Replay Attack Video Datasets



OUR OFFERING

- Off-the-shelf anti-spoofing video datasets
- High-quality metadata tagging

KEY STATS

25,000
total videos

(12,500 real videos, 12,500 replay attack videos)

12,500
unique
participants

5 ethnicity
groups
represented
in the dataset

Phased
delivery:
4 batches of
6,250
videos each

Metadata
attributes:
12 key
parameters
for enhanced
dataset usability

OVERVIEW

Shaip partnered with a leading AI security company to provide a high-quality, off-the-shelf anti-spoofing video dataset designed to enhance AI model training for fraud detection. The dataset included 25,000 videos capturing both real & replay attack scenarios, ensuring robust training data for anti-spoofing models.

Each of the **12,500 participants** contributed two videos - one real and one replay attack - recorded at **720p or higher resolution** with a frame rate of **26 FPS and above**.

The project's goal was to deliver **authentic and diverse datasets** that would enable AI models to effectively distinguish between real and spoofed biometric videos, thereby reducing fraud risks in biometric authentication systems.

PROJECT SCOPE

Dataset Curation

The project focused on delivering high-quality anti-spoofing video datasets consisting of **real and replay attack videos**. Key aspects included:

- » **12,500 participants** contributing **two videos each** (1 real, 1 spoofed).
- » **Diversity in recording devices** to enhance model adaptability.
- » **Balanced ethnic representation** to ensure dataset inclusivity.

Metadata Collection

Each video was accompanied by **12 metadata attributes** to enhance dataset usability.

CHALLENGES



- » **Ensuring Equal Representation:** Maintaining balanced ethnicity-wise data distribution while sourcing high-quality videos.



- » **Quality Control:** Ensuring that each participant contributes **one real & one replay attack video** to maintain dataset integrity.



- » **Technical Consistency:** Adhering to strict guidelines for **FPS (≥ 26), resolution ($\geq 720p$), and timestamp accuracy ($\pm 0.5ms$).**

SOLUTION

Shaip provided a structured and high-quality dataset to meet the client project's requirements. The solution included:

Dataset Curation & Quality Control

- » **25,000 videos** collected across **4 phases** to ensure a steady and structured data flow, avoiding bottlenecks.
- » **Rigorous validation process** to ensure compliance with **FPS, resolution, and metadata accuracy**. Each video underwent multiple quality checks before final acceptance.
- » **Comprehensive metadata tagging** with **12 attributes**:
 - File ID/Name
 - Type of Attack (Real/Replay)
 - Person ID
 - Video Resolution
 - Video Duration
 - Ethnicity of the Subject
 - Gender of the Subject
 - Whether Video is Original or Spoofed
 - Device Name/Model
 - Person Speaking or Not
 - Timestamp Start Time
 - Timestamp End Time
- » **Balanced Ethnic Group Distribution**: The dataset was meticulously curated to maintain a balanced ethnic representation. The distribution includes Hispanic (33%), South Asian (21%), Caucasian (20%), African (15%), and East Asian & Middle Eastern populations (each comprising up to 6%).
- » **No duplicate entries** to maintain dataset uniqueness and prevent biases in AI training.
- » **Ethnically diverse participant selection** to create a dataset that reflects real-world user variations, improving AI model adaptability and fairness.
- » **Recording device variation** included multiple smartphone models, cameras, and lighting conditions to enhance the model's robustness against different environmental settings.

THE OUTCOME

The high-quality, diverse anti-spoofing video dataset provided by Shaip enabled the client to **train AI models to accurately differentiate between real and spoofed videos** in various biometric authentication scenarios. The dataset contributed to:

- » **Improved Fraud Detection:** Enhanced AI performance in detecting fraudulent biometric attacks.
- » **Diverse Training Data:** Strengthened the model's ability to recognize replay attacks across different ethnicities, devices, and environmental conditions.
- » **Scalability:** The dataset serves as a foundation for future anti-spoofing model enhancements and expansions.

CUSTOMER TESTIMONIAL



"Shaip's dataset has been instrumental in enhancing our AI-driven anti-spoofing models. The diversity, quality, and structured metadata provided a strong foundation for improving fraud detection in biometric authentication systems."

ABOUT SHAIPI

Shaip provides high-quality data across multiple data types (text, audio, image & video) to companies looking to build unbiased and high quality AI/ML models. Shaip licenses, collects and annotates data for Healthcare, Conversational AI, Computer Vision and Generative AI/LLM use cases.

ABOUT THE CLIENT

Our client is a pioneer in AI security, specializing in biometric authentication and fraud prevention solutions for global enterprises. Their solutions safeguard digital transactions and enhance security in financial and enterprise applications.