AI-Powered Screen Recording & Auto Script Generation Platform

1. Project Overview

This project focuses on building a screen recording application that not only captures user activity but also intelligently converts recorded actions into a structured, step-by-step script. The system further enhances usability by generating a synchronized text-to-speech (TTS) video based on the recorded workflow.

The objective was to create a proof of concept capable of capturing screen interactions, interpreting user actions, and automatically generating instructional content without manual input.

2. Problem Statement

Traditional screen recording tools only capture video, requiring users to manually:

- Write step-by-step guides
- Annotate actions
- Create tutorials

This process is time-consuming and inefficient, especially for repetitive workflows or documentation tasks.

The challenge was to develop a system that can:

- Automatically detect user actions (clicks, scrolls, typing)
- Understand context (active window, UI elements)
- Generate meaningful step-by-step instructions
- Synchronize actions with video playback

3. Proposed Solution

A smart screen recording system that combines:

- Screen capture
- Event tracking
- OCR-based UI recognition
- Script generation
- Text-to-speech video creation

The system records user interactions and transforms them into:

1. A structured text script
2. A narrated step-by-step video

4. Core Features

Screen Recording

The application records the full screen or a selected region. Users can:

- Start recording with a countdown
- Pause, resume, and stop recording
- Record specific areas or the entire screen

The recording is saved in video format (MP4).

Action Tracking

The system captures all user interactions in real time, including:

- Mouse clicks
- Drag and drop actions
- Scroll events
- Keyboard inputs

Each action is recorded with:

- Timestamp
- Cursor position
- Event type

Active Window Detection

The application identifies which window the user is interacting with at any given time.

This allows the generated script to include contextual instructions such as:

- “On window XYZ, click Cancel”
- “Switch to window ABC and click Yes”

UI Element Recognition

To identify the labels of buttons (e.g., Cancel, Yes, No), the system:

- Captures the screen region around the click
- Applies OCR to extract text
- Maps the extracted text to user actions

This eliminates the need for manual script editing.

Script Generation

The system automatically generates a chronological script based on recorded actions.

Example output:

1. On window XYZ, click Cancel
2. Scroll down and click Yes
3. On window ABC, click No

The script is:

- Time-synchronized
- Editable by the user
- Exportable as a text file

Text-to-Speech Video Generation

The generated script is converted into a narrated video by:

- Mapping each step to its corresponding timestamp
- Generating voice narration using TTS
- Overlaying instructions on the recorded video

This results in a fully automated tutorial video.

Editor Module

After recording, users can refine the output using an editor.

Features include:

- Video playback
- Cropping sections
- Blurring sensitive areas
- Adjusting script text
- Language selection

Script Management

Users can:

- View generated scripts
- Edit instructions
- Export scripts to document formats
- Translate scripts into different languages

Recents and Sharing

The system maintains a history of recordings.

Users can:

- Access recent recordings
- Share videos via links
- Embed tutorials into external platforms

5. System Workflow

1. User starts screen recording
2. System captures:
    - Video stream
    - Mouse and keyboard events
    - Active window data
3. Each event is timestamped and stored
4. OCR extracts UI labels from clicked regions
5. Actions are converted into structured steps
6. Script is generated automatically
7. Script is synchronized with video timeline
8. TTS generates narration
9. Final instructional video is produced

6. Technology Stack

Core Technologies

- Python for automation and event tracking
- OpenCV for image processing
- OCR engines for text extraction
- Node.js (optional) for backend services

Libraries Used

- PyAutoGUI for mouse tracking
- Keyboard library for key events
- PyGetWindow for active window detection
- Tesseract OCR for UI text recognition

- FFmpeg for video processing
- Text-to-Speech engines for narration

7. Challenges and Solutions

Capturing Accurate User Actions

Different types of interactions needed precise tracking.
This was solved by combining mouse, keyboard, and system-level event listeners.

Identifying UI Elements

Buttons and labels are not directly accessible.
OCR was used to extract text from screen regions dynamically.

Synchronizing Script with Video

Ensuring accurate timing between actions and narration was critical.
Timestamps were recorded for every event and mapped to video frames.

Handling Multiple Windows

Users often switch between applications.
Active window tracking ensured correct context in the script.

Processing Performance

Real-time recording and processing required optimization.
Efficient event logging and post-processing pipelines were implemented.

8. Outcome

The system successfully demonstrated a working proof of concept that:

- Automatically converts screen activity into structured documentation
- Generates step-by-step instructional scripts
- Produces synchronized tutorial videos
- Eliminates the need for manual documentation

9. Future Enhancements

- AI-based UI understanding beyond OCR
- Browser extension version
- Cloud-based processing
- Collaboration features
- Advanced editing timeline
- Multi-language voice generation

10. Conclusion

This project introduces a new approach to documentation and tutorial creation by combining screen recording with intelligent automation. By transforming user actions into structured scripts and narrated videos, the system significantly reduces the effort required to create instructional content.

The solution is scalable and can be extended into a full product for training, onboarding, and knowledge sharing across industries.