Enhancing Art Perception for Individuals with Visual Impairments Using Computer Vision, Depth Perception and Multimodal Feedback

ADYA AGGARWAL, IIIT-Delhi
PANKHURI SINGH, IIIT-Delhi
PRANAV JAIN, IIIT-Delhi

Art remains largely inaccessible to individuals with visual impairments, limiting their ability to experience paintings. This paper presents a multimodal haptic-audio system that enables users to perceive artwork through touch and sound. A camera tracks a user's finger, marked with a distinct color, as it moves across a canvas, mapping its position to the artwork. A depth estimation algorithm translates image depth into haptic feedback using ERM vibration motors or piezoelectric actuators on the fingertips. Simultaneously, color information is converted into auditory cues via headphones. This dual-feedback mechanism enables vision-impaired users to experience spatial textures and color representations, allowing for a richer and more intuitive engagement with visual art. Additionally, sighted users gain a tactile appreciation of paintings that are typically untouchable, enriching their interaction with art.

Additional Key Words and Phrases: Multisensory Art Accessibility, Haptic Feedback for Art, Computer Vision, Depth Mapping for Tactile Art, Color Sonification

ACM Reference Format:
Adya Aggarwal, Pankhuri Singh, and Pranav Jain. 2018. Enhancing Art Perception for Individuals with Visual Impairments Using Computer Vision, Depth Perception and Multimodal Feedback. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 20 pages. https://doi.org/XXXXXXX.XXXXXXX

The complete source code is available on GitHub.

1 Introduction

Traditional two-dimensional (2D) paintings and artworks rely almost entirely on visual perception, making them inaccessible to individuals with visual impairments and limiting tactile engagement for sighted users. While some accessibility solutions exist, they often fail to provide a comprehensive sensory experience. Bas-relief techniques and tactile graphics attempt to translate paintings into touchable forms, but they frequently lose fine details and spatial accuracy. Audio descriptions, though helpful, often provide subjective interpretations rather than a structured understanding of an artwork's composition.

A major challenge in making visual art more accessible is the inability to convey depth, texture, and color in a way that is intuitive for non-visual users. Color, in particular, remains difficult to represent through tactile or auditory means, leaving individuals with an incomplete understanding of an artwork's visual elements. Additionally, existing accessibility tools lack interactivity, preventing users from actively exploring paintings at their own pace. These limitations create a gap in how visually impaired individuals and even sighted users interact with and appreciate visual art, highlighting the need for a more effective and inclusive solution.

To address these limitations, this research proposes an AI-driven, multimodal system that enhances art perception through computer vision, haptic feedback, and auditory cues.

Authors' Contact Information: Adya Aggarwal, IIIT-Delhi, New Delhi; Pankhuri Singh, IIIT-Delhi, New Delhi; Pranav Jain, IIIT-Delhi, New Delhi.

2018. Manuscript submitted to ACM

Problem Statement. Despite advancements in accessibility tools, there remains a lack of affordable, interactive, and intuitive systems for experiencing visual art without sight. Current solutions are either expensive or cognitively demanding, often sacrificing detail or interactivity. There is a need for a cost-efficient and engaging approach that can simulate depth and color perception while minimizing cognitive overload for the user.

Research Questions. To tackle the above challenge, this work explores the following key research questions:

(1) **How can we design a cost-effective haptic feedback system using basic vibration motors to simulate texture for visually impaired users?**
(2) **How can we reduce cognitive overload during art exploration to enhance the learning curve and ensure sustained user engagement?**
(3) **Can Large Language Models (LLMs) be integrated to provide contextual and adaptive learning experiences for users interacting with artworks?**

2 Literature Review

Accessibility in visual perception extends beyond the realm of art, influencing various fields such as sports, performance arts, music, and education. Researchers have explored different technologies and methodologies to enhance sensory experiences for individuals with visual impairments or other sensory limitations. Studies in haptic feedback, auditory substitution, computer vision, and multisensory integration have contributed to making information more accessible across diverse domains. The following sections review literature from multiple disciplines, highlighting advancements, limitations, and potential applications relevant to creating a more inclusive and immersive experience.

2.1 Haptic Assistance in Sports

A drone-guided haptic system helps visually impaired runners by sending directional cues via a wearable vibrotactile device. Continuous feedback proved most effective, keeping users on track 93% of the time while increasing their confidence and safety. Future developments aim to support visually impaired athletes in team sports like blind soccer[1].

A skiing assistance system uses vibrating bracelets on the forearms to provide haptic guidance. Ski instructors control these vibrations via augmented ski poles, allowing users to navigate without relying solely on auditory cues. This system enhances safety and reduces cognitive load for visually impaired skiers[2].

2.2 Multisensory Communication in Performing Arts

A vibrating bracelet enhances communication for deaf performers by providing tactile feedback linked to emotional states. Using heart rate monitoring, it helps actors understand and express emotions, improving audience engagement. The study highlights the role of multisensory technology in making the arts more inclusive[3].

2.3 Assistive Technologies for Deafblind Individuals

Deafblind individuals lack both visual and auditory channels for communication, making interaction difficult. Assistive devices use haptic stimuli to deliver textual information and translate speech and visual input into tactile feedback. These systems expand access to digital platforms and support independent communication[4].

2.4 Music Accessibility Through Multisensory Integration

The Auris System enables music experiences for individuals with hearing impairments by converting sound into tactile and visual stimuli. Devices like the Auris Chair and Bracelet translate musical elements into vibrations, activating the auditory cortex through multisensory integration. EEG analysis helps refine these representations, improving accessibility for deaf users[5].

2.5 Wearable Safety and Alert Systems

The Vibration Alert Bracelet provides emergency notifications for visually and hearing-impaired users through wireless alerts and distress signals. It features a Wi-Fi-enabled bracelet and a mobile app, offering a lightweight and efficient solution. Future upgrades may include BLE for power efficiency, GPS tracking, and enhanced security[6].

2.6 Crossmodal Translation Between Sound and Vision

Innovative approaches in sensory substitution explore the translation between auditory and visual modalities. One such method involves converting sounds into visual patterns or vice versa to support individuals with sensory impairments. Research highlights systems that use visual displays to represent auditory properties such as pitch, timbre, and rhythm, enabling deaf users to perceive music visually. Conversely, images can be sonified to convey visual scenes to blind users through sound patterns. These techniques leverage the brain's neuroplasticity and underscore its ability to interpret sensory data across modalities. The development of such systems reveals potential for inclusive media, education, and multisensory artistic performances[20]

2.7 Color-Based Quantification and Multisensory Translation

The integration of color-based representation into analytical systems extends beyond scientific measurement and into broader accessibility. Cantrell et al. introduced the use of the hue (H) parameter from the HSV color space as a stable and precise metric for bitonal optical sensors. Unlike RGB intensity-based approaches, the hue value remains consistent across variations in membrane thickness, indicator concentration, and imaging conditions. This makes it a reliable parameter for portable or disposable sensors used in mobile diagnostics and quality control. The study demonstrates how a traditionally qualitative visual attribute—color—can be quantified and used analytically, bridging the gap between visual perception and data-driven analysis[21]

2.8 Low-Cost Haptic Glove Designs

Recent advancements have led to the development of affordable haptic gloves aimed at enhancing accessibility and immersion in various applications.

DOGlove is an open-source glove offering 21 degrees of freedom (DoF) motion capture and 5-DoF force feedback, assembled under $600, designed for precise teleoperation and dexterous manipulation.

Pulse Haptic Glove, designed for enthusiasts, uses knuckle-mounted motors to apply pressure via fingertip caps, providing realistic haptic feedback at approximately $300.

TactGlove features 10 individually controllable linear resonant actuators (LRAs) on the fingertips, offering enhanced haptic feedback compatible with camera-based hand tracking systems, priced around $299.

Affordable haptic gloves beyond the fingertips propose a novel design extending haptic feedback beyond fingertips to include intermediate and proximal phalanges, utilizing a ratchet and pawl mechanism for more realistic feedback.

3 Methodology

The methodology is structured into four key phases: data collection, system pipeline design, code development, and hardware integration.

3.1 Data Collection

Multiple data collection methods were employed to ensure a user-centered design and effective system development. These included interviews, surveys, and persona building, providing both qualitative and quantitative insights.

3.2 System Pipeline and Workflow

This section elaborates on the technical pipeline driving the system. It integrates computer vision, depth estimation, color classification, and multisensory feedback to create an immersive experience.

3.2.1 ***Image and Surface Setup.*** A printed artwork is placed on a flat surface. A camera mounted overhead captures the entire frame including the image and the user's hand. This is done by recognizing the image by a big green rectangle.

3.2.2 ***Depth Map Generation.*** The image is processed via the SculptOK depth estimation API, which returns a pixel-wise depth map using pre-trained models. This map simulates surface texture for tactile conversion.

3.2.3 ***Real-Time Hand and Finger Tracking.*** Using OpenCV and MediaPipe Hands, the system detects and continuously tracks the index fingertip (Landmark ID 8). The fingertip's (x, y) coordinates are mapped to the image frame and corresponding depth data.

3.2.4 ***Vibration Feedback (Depth-to-Haptics).***

$$I(d) = \begin{cases} 0 & \text{if } d \leq 0.01 \\ d^{1.5} \cdot \left(1 - (1-d)^{1.5}\right) & \text{if } d > 0.01 \end{cases}$$

Where:

- $d \in [0, 1]$ is the *normalized depth* value (i.e., depth_map$[y, x]/255.0$),
- $I(d)$ is the *vibration intensity*.

This function produces a *bell-shaped curve* peaking at intermediate depth values.

3.2.5 ***Color Detection and Audio Feedback.*** The RGB color at the fingertip location is classified using a trained K-Nearest Neighbors (KNN) classifier. The following mappings are used:

- **Red**: Guitar
- **Black**: Saxophone
- **Blue**: Flute
- **Green**: Congo
- **Yellow**: Piano

These are sonified and played using the Python `sounddevice` library. All sounds are defined by 3 important metrics - Frequency(Hz), Amplitude and Waveform. The waveforms have been modified to simulate different instruments, with a frequency of 440Hz and equal amplitu.

3.2.6 ***Integrated Feedback Experience***. As the user moves their finger across the artwork:

- Texture (depth) is conveyed via vibration intensity.
- Color is conveyed through musical audio cues.

This creates a synchronized haptic-audio experience for visually impaired users to explore visual art through touch and sound.

3.3 Hardware Integration

The system relies on a combination of software and hardware to deliver synchronized feedback. Below is an overview of the hardware connections and purpose:

*1. Arduino UNO Acts as the interface between depth-based intensity and the ERM motor. Receives vibration levels via serial input from the main system.

*2. ERM Vibration Motor Connected to a PWM-enabled digital pin (e.g., D9) on the Arduino. Provides real-time feedback at varying intensities.

*3. Camera Mounted overhead, continuously streams frames to the Python script for fingertip detection and position mapping.

*4. Headphones Connected to the system output. Sound cues are generated on the host machine and played back instantly to indicate color recognition.

*5. Power Supply The Arduino is powered via USB from the main system, and external power may be used for consistent motor strength if required.

This tight integration allows the user to interact naturally with an image and receive immediate sensory feedback, enabling an intuitive and accessible form of art exploration.

3.4 Chatbot Integration

In the language model pipeline, the process begins with an image captured during the initial phase, accompanied by a natural language question provided by the user. This image-question pair is first processed by BLIP-2, a state-of-the-art Vision-Language model designed for Visual Question Answering (VQA). BLIP-2 interprets the visual content of the image in conjunction with the user's query and generates a relevant textual response based on its understanding of both modalities.

The output from BLIP-2 is then passed to Mistral-7B, a powerful open-weight large language model known for its instruction-following and text-generation capabilities. Mistral-7B takes the VQA response and contextualizes it further, refining the output into a coherent, informative, and well-articulated sentence. This final response combines the visual insight from BLIP-2 with Mistral-7B's linguistic fluency, enabling a more conversational and context-aware interaction.

3.5 Modalities

The system employs a multimodal interaction framework designed to reduce cognitive load and enhance efficiency by distributing information across various sensory channels. This approach aligns with principles of multisensory integration, where combining inputs from different modalities can lead to improved perception and performance.

In the primary system, users explore digital artworks through tactile and auditory feedback. A camera tracks the user's finger movements across a canvas, mapping these to corresponding elements in the artwork. Depth information is conveyed through haptic feedback using Eccentric Rotating Mass (ERM) vibration motors on the fingertips, allowing

users to perceive spatial structures. Simultaneously, color information is translated into distinct auditory cues delivered via headphones, enabling users to identify colors through sound.

Complementing this, a conversational chatbot interface supports both text and speech interactions. Users can input queries either by typing or speaking, and the chatbot responds with both textual and spoken answers. This dual-mode communication caters to different user preferences and situational needs, facilitating seamless interaction.

By integrating tactile, auditory, and linguistic modalities, the system distributes cognitive demands, allowing users to process information more naturally and efficiently. This multimodal design not only enhances user engagement but also supports accessibility, particularly for individuals with visual impairments.

4 Evaluation and Explaination

The project code is modularized into the following key components:

4.1 Explainability

To enhance the transparency and accountability of our model's predictions, we applied **SHAP (Shapley Additive exPlanations)**, a leading explainability framework grounded in cooperative game theory. SHAP provides instance-level attributions that clearly explain how much each input feature (Red, Green, and Blue channel intensities) contributed to the model's final prediction.

The following table displays SHAP values along with the raw RGB inputs and the predicted color label for some samples:

Index	R	G	B	SHAP_R	SHAP_G	SHAP_B	Predicted_Label
0	4	222	10	0.077	**-0.293**	0.117	Green
1	19	17	234	0.111	0.083	**-0.294**	Blue
2	223	35	1	**-0.270**	0.050	0.120	Red
3	48	243	32	0.071	**-0.293**	0.122	Green
4	227	34	39	**-0.270**	0.050	0.120	Red

Table 1. SHAP Empowers Confident Interpretation

Key Observations

(1) The model consistently predicts correct classes despite complex input combinations. For instance, in Sample 0, even though the Green (G) channel shows a strongly negative SHAP value, the model balances it with positive contributions from the Red and Blue channels to confidently classify it as "Green". This shows a mature decision boundary, where the model does not rely on a single dominant feature but understands color nuances holistically.

(2) In Sample 1, the model predicts "Blue" despite the Blue (B) channel having a negative SHAP value (-0.294). This again proves that the model has learned fine-grained relationships, where high blue intensity alone does not

blindly drive the decision — instead, it factors in the subtle push from the Red and Green channels for a more context-aware classification.

(3) The Red class predictions in Samples 2 and 4 reflect that high Red intensity does not necessarily mean a strong SHAP$_R$*value*.*Interestingly*, *bothhaveslightlynegativeSHAP*$_R$*contributions*, *meaningthemodelislearningtodifferentiatetrueredsfromredswi dimensionalways*.

4.2 Model Evaluation: Hand Recognition System Using MediaPipe

To evaluate our hand recognition pipeline, we tested it on the HaGRID 30K sample dataset (384p resolution). The model leverages MediaPipe Hands, a lightweight and highly optimized framework by Google that performs real-time hand tracking using a machine learning model trained to detect up to 21 3D landmarks per hand. The detection performance is summarized below:

(1) **Precision(1.000)** : This indicates that every detected hand was a true positive—there were no false positives. MediaPipe's confidence-based detection and landmark refinement pipeline ensures that once a hand is identified, it is highly reliable and accurate.

(2) **Recall (0.674)**: While MediaPipe is precise, it misses about 32.6% of the hands in the dataset. This could be due to:occlusions,multiple hands in frame,low lighting or resolution issues,MediaPipe's internal thresholding (designed for speed over exhaustive detection).

(3) In a high-variance dataset like HaGRID, which includes diverse poses and gestures, this trade-off is expected.

(4) **F1 Score (0.805):** A harmonic mean of precision and recall, the F1 Score reflects an overall balanced performance. While recall could be improved with more aggressive detection tuning or ensemble methods, the score indicates a strong and stable performance.

(5) **Accuracy (0.674):** This metric shows the overall percentage of correct detections (true positives + true negatives). Since this is a detection task, accuracy here is consistent with recall, reaffirming that most misses are due to undetected hands, not incorrect ones.

Why MediaPipe Works Well :

(1) **Real-time optimized:** Extremely efficient on CPU and mobile.

(2) **Precision-first:** Prioritizes correctness over exhaustive detection, as reflected in the high precision.

(3) **Modular:** Can be extended with custom classifiers (e.g., gesture recognition, hand sign classification).

Our MediaPipe-based hand recognition system demonstrates excellent precision and reliable performance, particularly suitable for real-time applications where false positives must be minimized. With additional tuning or post-processing, recall can be improved to further enhance detection coverage without sacrificing accuracy.

Performance of functions used in contour detection

(1) **cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)**: Converts a BGR image to grayscale. On CUDA-enabled GPUs, the device execution time is approximately 0.03 ms, with host times ranging from 0.087 ms to 0.28 ms, depending on system load and initialization overhead.

(2) **cv2.GaussianBlur(gray, (5, 5), 0)**: Applies Gaussian blur to reduce image noise and detail.Ikomia +15. Virtual Edge Solution | IO River +15.BoofCV +15.On a GeForce GTX 750 GPU.

(3) **cv2.Canny(blur, 50, 150)**: Performs Canny edge detection.The execution time varies with image complexity. For instance, on an 8-threaded system, optimization reduced the Canny algorithm's runtime from 1.7 seconds to 1.45 seconds.

4.3 Model Evaluation: Color Identification Using KNN

The K-Nearest Neighbors (KNN) classifier used for color identification demonstrates strong predictive performance, as reflected in the following metrics:

(1) **F1 Score (0.9715):** This high F1 score highlights that the model maintains an excellent balance between precision and recall, meaning it is effective at correctly identifying colors while minimizing both false positives and false negatives. This is particularly important in color classification, where overlaps between visually similar colors can occur.
(2) **Accuracy (0.9720):** An accuracy of 97.2% confirms that the model classifies the vast majority of test samples correctly. It suggests that the current feature set (RGB values) and KNN's distance-based classification are highly effective for this task.

With both high accuracy and a strong F1 score, the KNN model proves to be a robust and reliable choice for color identification. Its performance can serve as a solid baseline, with opportunities to further optimize through parameter adjustments.

4.4 Chatbot Evaluation

:

1. Speech-to-Text (STT) – Google Cloud API

(1) **Word Error Rate (WER):** Approximately 8% under favorable conditions.
(2) **Latency:** Typically 1–2 seconds for short audio inputs.
(3) **Multilingual Support:** Supports over 125 languages, including English and Hindi.

2.Text-to-Speech (TTS) – ElevenLabs

(1) **Mean Opinion Score (MOS):** Ranging from 4.2 to 4.5 on a 5-point scale, indicating high naturalness.
(2) **Latency:** Approximately 5–6 seconds for generating short responses.
(3) **Features:** Offers emotional tone control, multilingual support, and real-time streaming capabilities.

3.Chatbot Evaluation (Visual Question Answering)

(1) **BLEU Score:** Typically ranging from 0.25 to 0.35, reflecting moderate alignment with reference answers.

4.5 Interpretability

The Odin Vision system is explainable and interpretable because it uses transparent, human-understandable logic to convert visual elements like color and depth into sound and vibration, allowing visually impaired users to meaningfully interact with paintings. Its modular design, reliance on simple models like KNN for color classification, and real-time visual overlays ensure the decision-making process is easy to understand, debug, and explain—aligning with the course definitions of explainability and interpretability. By mapping distinct colors to culturally intuitive sounds and varying vibration based on proximity, the system offers consistent, predictable feedback that fosters trust, usability, and inclusivity. Additionally, features like voice-based interaction and multilingual support make it accessible to diverse users, supporting the principles of human-centered AI, including transparency, user control, and cultural sensitivity.

5 User Study and Evaluation

Fig. 1. Participant 1

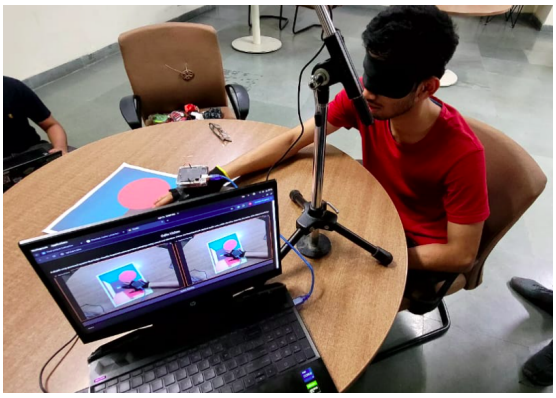

Fig. 2. Participant 2

5.1 Pre-Study Survey

To evaluate the effectiveness of our system designed for visually impaired individuals, we conducted a study involving 10 participants who were blindfolded for the duration of the experiment. Before interacting with the system, each participant was asked to fill out a preliminary survey. This survey included questions about any pre-existing visual impairments. The majority of participants reported no such impairments. When asked about their engagement with visual arts, 27 participants stated they engage frequently, 36 occasionally, 27 rarely, and 9 never. On being asked whether they had ever experienced art through non-visual means, 63% responded negatively. Most participants rated their sense of touch for object recognition as a 3 on a scale of 1 to 5.

5.2 Study

The core system used in this study tracks a user's finger via a camera, mapping its position across a canvas in correspondence with the artwork. A depth estimation algorithm translates image depth into haptic feedback using ERM vibration motors attached to the fingertips. Simultaneously, color information from the artwork is converted into auditory cues delivered through headphones. The task for participants was to experience the painting through this system and then attempt to recreate it by drawing.

5.3 Compositional Variation

The images that participants drew displayed consistent variation, yet key elements from the original artwork were present. These included a large red circle against a blue background, boxes of various colors in the foreground, and a green triangle set against a yellow background. In terms of shape recognition, most participants were able to identify the basic geometric forms; however, they faced challenges with accurate composition, particularly when visual elements were overlapping.

5.4 Colour Variation

Color identification showed promising results. Most participants correctly identified the colors used in the artwork. Since each color was paired with a distinct auditory cue, recognition was relatively successful. However, the outline of the shapes remained difficult for participants to reproduce accurately. While the locations where the colors were applied were mostly correct, the shapes themselves were often distorted or imprecise.

5.5 Post-Study Reflection

In the post-study reflection, 45% of the participants rated the difficulty of understanding the artwork through touch and sound as 3 out of 5, and another 45% rated it as 4. A majority of participants (63%) found the vibration feedback to be somewhat clear, and 45% reported that the auditory cues were somewhat helpful in identifying colors. While the system showed potential, participants noted areas for improvement. Suggestions included enhancing the language model's responses to give more informative guidance, and improving the comfort of the haptic glove. Notably, 72% of the participants expressed interest in using such a system in real-world settings like museums and galleries.

Fig. 3. Participant recreations of the artwork based on multisensory interaction

5.6 Conclusion

Overall, the study highlighted both the promise and the limitations of the system. While participants could recognize and reproduce fundamental aspects of the artwork through multisensory cues, challenges remain in achieving precision in shape and composition. Nonetheless, the positive reception suggests strong potential for further development and application of such assistive technologies in inclusive art experiences.

6 User Evaluation HCI Principles

Tactile Feedback and Affordance

The glove integrates a rubber-dot textured surface to enhance tactile interaction, supporting the principle of *affordance*. The textured dots offer clear interaction cues, enabling users to understand how to interact with the system based

on surface feedback. This tactile design aligns with HCI's goal of reducing cognitive load and ensuring intuitive user interaction by providing immediate, localized haptic feedback.

Cognitive Load Optimization

Initially, the system's mapping of *Hue*, *Saturation*, and *Value* to *Instrument*, *Frequency*, and *Amplitude*, respectively, was found to be cognitively demanding. This violated the HCI principles of simplicity and error prevention. In response, the system was redesigned to map only **Hue** to **Instrument**, reducing the cognitive load and supporting easier recognition over recall. This change also adheres to the principle of *minimizing cognitive load*, making the interaction process more intuitive and user-friendly.

Depth-Based Object Perception

To enhance the user experience, the system utilizes depth mapping to aid object recognition. By providing haptic feedback based on the depth of objects within a scene, users can perceive the spatial arrangement of these objects. This design follows the HCI principles of *spatial awareness* and *natural interaction*, ensuring users can intuitively navigate and interact with the environment in a way that is both rewarding and engaging.

Inclusive Dataset Utilization

The system's use of the HaGRID dataset, which includes diverse skin tones and hand types, demonstrates a commitment to *inclusive design*. By ensuring the model's robustness across varied demographics, the system minimizes bias and provides consistent interaction experiences for all users. This supports the HCI principle of *designing for all*, ensuring the technology is accessible and effective for a wide range of users.

Multimodal Interaction via Voice Interfaces

The system employs voice interfaces, utilizing **Google Speech-to-Text** optimized for Indian accents and **ElevenLabs Text-to-Speech** using Indian voice models. This integration enhances the system's *accessibility* and *engagement* by providing a familiar, culturally inclusive interface. By facilitating natural voice-based interaction, the system fosters a more sociable and intuitive user experience, aligning with HCI goals of reducing barriers and promoting seamless communication.

User Feedback Loop

At the end of each interaction, the system actively gathers user feedback. This feedback loop embodies *user-centered design*, where user insights directly inform future design iterations. Encouraging ongoing user involvement aligns with HCI's emphasis on *continuous improvement*, building trust and reinforcing the value of the user's perspective in shaping the system.

Error Handling and System Status Visibility. The system includes real-time error notifications and progress indicators, offering users clear visibility into the system's status. This transparency aligns with the HCI principles of *visibility of system status* and *feedback*, ensuring users are always informed and never left uncertain about the system's operations. Clear error handling and status visibility minimize frustration, fostering a positive user experience and increasing trust in the system.

7 Discussion

7.1 Low-Resource Haptic Glove

Our system utilizes a cost-effective haptic glove built with ERM vibration motors controlled via an Arduino UNO. This glove offers five discrete levels of vibration based on image depth, allowing users to perceive texture variations. Although more advanced gloves such as DOGlove or TactGlove offer greater precision and multimodal feedback, our implementation prioritizes accessibility and affordability. It can be easily deployed in resource-constrained environments like schools or community spaces, enabling a broader impact.

7.2 LLM as a Guide: Chatbot for Knowledge Validation

We propose the integration of a Large Language Model (LLM) to act as an intelligent guide or chatbot within the system. Beyond tactile and auditory feedback, the LLM would provide real-time explanations, historical context, and conversational validation for user understanding. For example, after identifying a region of the artwork, users could inquire about its symbolism, artist, or era. The LLM could also quiz the user to validate their comprehension, facilitating an engaging and educational interaction—particularly valuable for visually impaired users who rely heavily on descriptive interpretation.

8 Limitations

While the proposed system shows promise, several limitations need to be acknowledged:

- **Basic Vibration Feedback:** The use of ERM motors limits the system's ability to convey subtle tactile differences, affecting the perception of intricate textures and gradients.
- **Resource Constraints:** The focus on affordability restricts the use of higher-end sensors, actuators, and computing resources that could enhance the fidelity of the experience.
- **Limited Participation from Visually Impaired Individuals:** Due to access constraints, most evaluations were conducted with sighted users simulating visual impairment. This may not reflect true usability for the intended demographic.
- **Color-to-Sound Mapping Simplification:** The mapping of six colors to six instruments simplifies interpretation but fails to represent the full diversity of hues and tonal variations found in art.
- **Environmental Sensitivity:** The system's color recognition is sensitive to lighting changes, which can impact the accuracy of auditory feedback.

9 Future Work

Several extensions and improvements are envisioned for future iterations of the system:

- **Higher-Quality Haptics:** Incorporation of Linear Resonant Actuators (LRAs) or piezoelectric feedback mechanisms could offer more realistic and localized haptic responses.
- **Complex Image Interpretation:** Future versions will aim to handle high-resolution artworks, segment artistic elements, and identify stylistic features using advanced computer vision.
- **Color Gradation to Soundscape:** Instead of mapping discrete colors to instruments, continuous color gradients could be translated into ambient soundscapes using variations in pitch, tone, and reverb.
- **Emotion-Based Sonification:** The system could identify the emotional tone of artwork and generate corresponding music to enhance interpretive immersion.

- **Expanded Multimodal Interfaces:** Adding voice narrations, spatial audio, or even tactile displays could improve inclusivity and enrich the user experience.
- **Collaborative Exploration:** Introducing a multi-user mode could facilitate group learning and enhance social engagement with the artwork.

References

[1] Boreas Technologies. *Guidelines of Haptic UX Design*. Retrieved from https://pages.boreas.ca/blog/piezo-haptics/guidelines-of-haptic-ux-design, accessed April 2025.

[1] . Liao, J. Salazar, and Y. Hirata, "Robotic Guidance System for Visually Impaired Users Running Outdoors Using Haptic Feedback," 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 8325–8331, Sep. 2021, doi: https://doi.org/10.1109/iros51168.2021.9636567.

[2] . Aggravi, G. Salvietti, and D. Prattichizzo, "(PDF) Haptic Assistive Bracelets for Blind Skier Guidance," www.researchgate.net, Jan. 2016. https://www.researchgate.net/publication/292681848$_{H}aptic_{A}ssistive_{B}racelets_{f}or_{B}lind_{S}kier_{G}uidance(accessedMar.22,2022)$.

[3] . H. Huang and Y. Li, "The Development and Application of the Wearable Device for the Deaf Performers," ISSN: 2189-101X – The Asian Conference on Education International Development 2023 Official Conference Proceedings, pp. 579–591, May 2023, Accessed: Feb. 20, 2025. [Online]. Available: https://papers.iafor.org/submission68872/

[4] . Ranasinghe et al., "EnPower: Haptic Interfaces for Deafblind Individuals to Interact, Communicate, and Entertain," Advances in Intelligent Systems and Computing, pp. 740–756, Nov. 2020, doi: https://doi.org/10.1007/978-3-030-63089-8$_{4}$9.

[5] . Alves Araujo, F. Lima Brasil, A. Candido Lima Santos, L. de Sousa Batista Junior, S. Pereira Fonseca Dutra, and C. Eduardo Coelho Freire Batista, "Auris System: Providing Vibrotactile Feedback for Hearing Impaired Population," BioMed Research International, vol. 2017, pp. 1–9, 2017, doi: https://doi.org/10.1155/2017/2181380.

[6] . Conley et al., "Vibration Alert Bracelet for Notification of the Visually and Hearing Impaired," Journal of Open Hardware, Oct. 2019, Available: https://epublications.marquette.edu/electric$_{f}ac/622/$

[8] Q. Zeng, R. R. Martin, L. Wang, J. A. Quinn, Y. Sun, and C. Tu, "Region-Based Bas-Relief Generation from a Single Image," *Graphical Models*, vol. 76, no. 3, pp. 140-151, 2014.

[9] X. Sun, P. L. Rosin, R. R. Martin, and F. C. Langbein, "Bas-Relief Generation Using Adaptive Histogram Equalization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 4, pp. 642-653, 2009.

[10] J. Kerber, A. Tevs, A. Belyaev, R. Zayer, and H.-P. Seidel, "Feature-Sensitive Bas-Relief Generation," *IEEE International Conference on Shape Modeling and Applications (SMI)*, 2009.

[11] X. Wu, Y. Zhao, J. Luo, M. Zhang, and W. Yang, "Bas-Relief Modeling from RGB Monocular Images with Regional Division Characteristics," *Scientific Reports*, vol. 12, 2022.

[12] N. Suciati, M. T. Baihaqi, H. Tjandrasa, A. Z. Arifin, et al., "Converting Image into Bas Reliefs Using Image Processing Techniques," *Journal of Physics: Conference Series*, vol. 1196, 2019.

[13] Y. Fu, H. Yu, C.-K. Yeh, J. Zhang, and T.-Y. Lee, "High Relief from Brush Painting," *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[14] Z. Yang, B. Chen, Y. Zheng, X. Chen, and K. Zhou, "Human Bas-Relief Generation from a Single Photograph," *IEEE Transactions*, 2021.

[15] P. Sobociński, M. Maik, and K. Walczak, "Multimodal Presentation of 3D Relief Sculptures in Virtual Reality," *International Conference on Cyberworlds (CW)*, 2022.

[16] S. Kratz and T. Dunnigan, "ThermoTouch: Design of a High Dynamic Temperature Range Thermal Haptic Display," *CHI Extended Abstracts*, 2016.

[17] Thulin, S., "ThermoTouch: Sound maps matter: expanding cartophony. *Social & Cultural Geography*, *19*(2), 192–210. https://doi.org/10.1080/14649365.2016.1266028

[18] H. Ö. Sertlek, H. Slabbekoorn, C. ten Cate, and M. A. Ainslie, "Source specific sound mapping: Spatial, temporal and spectral distribution of sound in the Dutch North Sea," *Environmental Pollution*.

[19] W. Lin, "The hearing, the mapping, and the Web: Investigating emerging online sound mapping practices," *Landscape and Urban Planning*, vol. 142, 2015.

[20] Cantrell, K., Erenas, M. M., de Orbe-Payá, I., & Capitán-Vallvey, L. F. (2010). Use of the hue parameter of the hue, saturation, value color space as a quantitative analytical parameter for bitonal optical sensors. *Analytical Chemistry*, 82(2), 531–542. https://doi.org/10.1021/ac901753c

[21] Yakhontova, E. N. (2016). Translating sounds into visual images, and vice versa. *Psychology and Psychotechnics*, (6), 103–108. https://cyberleninka.ru/article/n/translating-sounds-into-visual-images-and-vice-versa

10 Appendix

10.1 Low-Fidelity and System Design

A low-fidelity prototype was developed using basic materials to simulate system flow. The user interface consists of:

- A static image placed on a table.

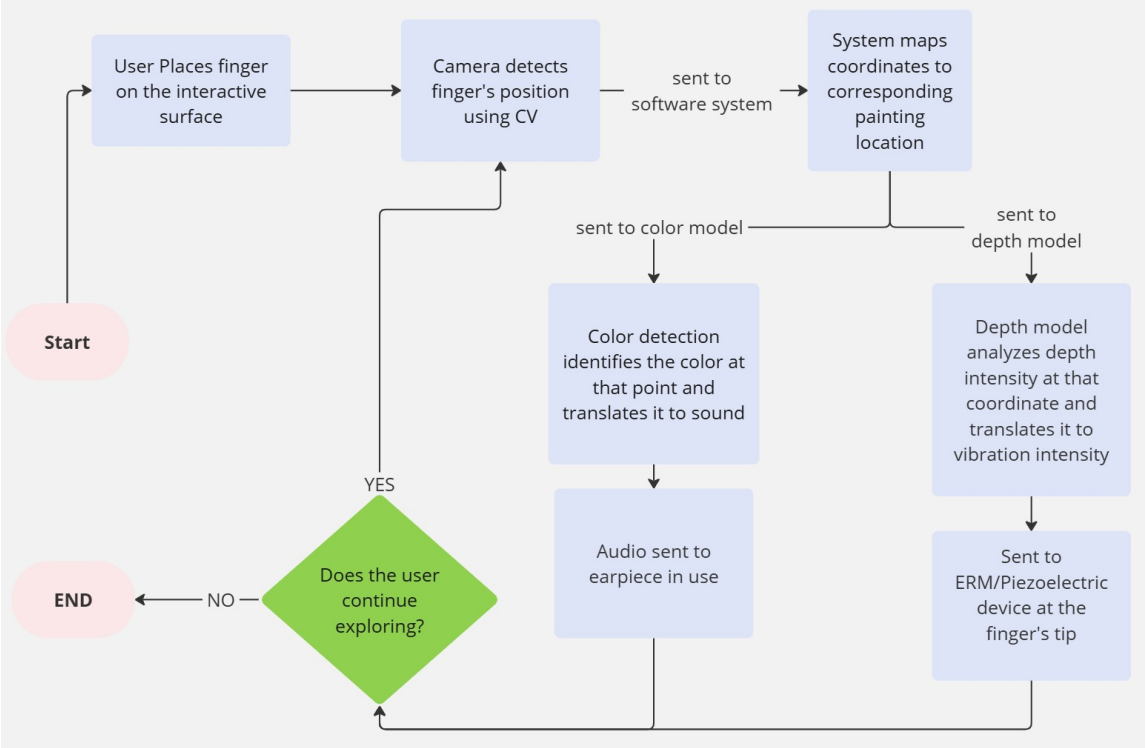

Fig. 4. User Taskflow

- A webcam overhead to track hand movement.
- Colorful fingertip markers to improve tracking reliability.
- Vibrating motors and a headphone setup to deliver multisensory feedback.

The system design follows a camera–processing–feedback loop, where all tracking and classification occur in real-time. This helps in validating the interaction flow and user response before progressing to final hardware implementation.

10.2 Low Fidelity Prototype

In Figure 1, the cursor, controlled by the fingertip, provides both haptic and auditory feedback based on its position within the visual field. When the cursor moves over darker regions, the vibration intensity remains low, and the corresponding sound frequency in the earphones is also lower. As the cursor moves into brighter, white regions, the vibration intensity increases, and the sound frequency rises accordingly. This combination of tactile and auditory feedback allows for an enhanced perception of different colors in the virtual environment

10.2.1 User Taskflow. See Figure 2.

10.2.2 Overview. Figure 3. illustrates an interactive system where a camera captures the user's hand movements, allowing them to explore a virtual interface using their fingertip as a cursor. As the user moves their finger, they

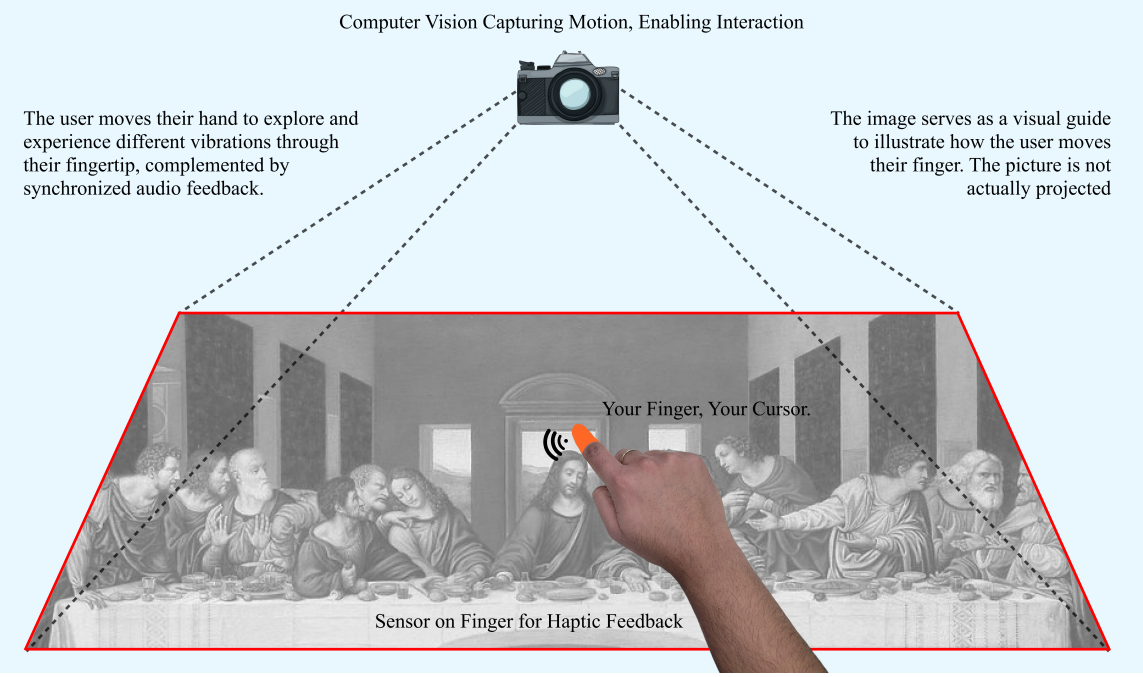

Fig. 5. Low Fidelity

experience different vibrations through a sensor attached to their fingertip, providing haptic feedback synchronized with audio cues. The displayed image serves only as a visual guide and is not actually projected.

10.3 Problems Solved

- **Limited Accessibility to Art for the Visually Impaired** : Traditional paintings are entirely visual, making them inaccessible to blind and visually impaired individuals. Existing solutions, like audio descriptions, fail to provide a fully immersive and interactive experience.
- **Lack of Tactile Engagement for Sighted Users**: Museums and galleries prohibit touching artwork, preventing visitors from experiencing the texture and fine details of paintings, limiting engagement and appreciation.
- **Insufficient Detail in Bas-Relief Generation**: Current 2D-to-bas-relief conversion methods often lack precision in preserving fine textures, focusing only on general depth rather than intricate details.
- **Absence of a Unified Multisensory Approach**: Previous research has addressed depth mapping, color sonification, and haptic feedback separately, but no comprehensive system integrates all three for a richer, more immersive artistic experience.

10.4 Mid Fi Design

This section presents the mid-fidelity prototype of the system, illustrating the functional flow and interactions between different components.

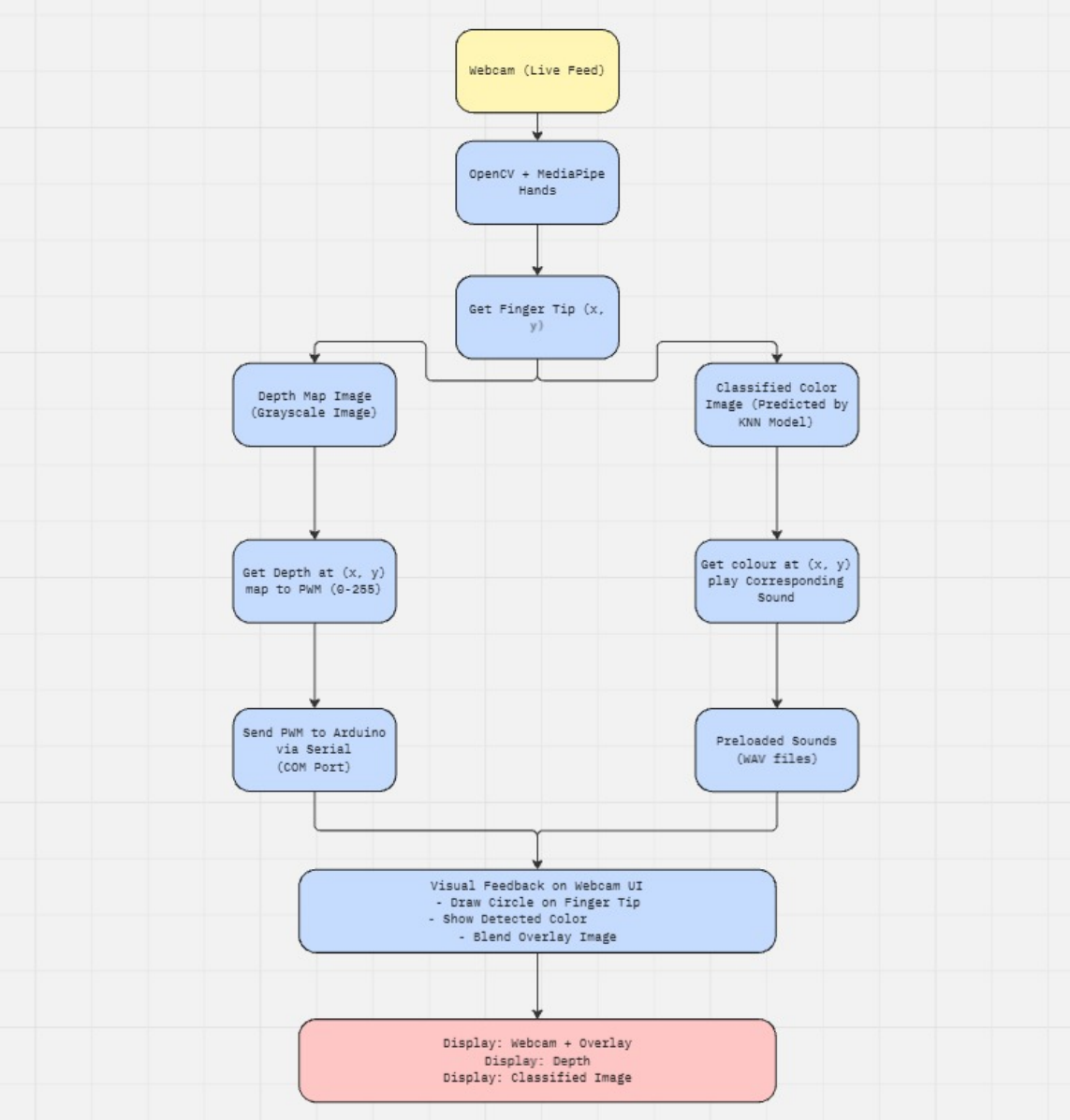

Fig. 6. TaskFlow

10.4.1 Flow Diagram. The flow diagram (see Fig 4)provides a structured representation of how the system processes input from the user, including depth mapping, color classification, and sensory feedback mechanisms.

10.4.2 Circuit Diagram. This circuit diagram (see Fig 5)illustrates the connection between an Arduino Uno (U1) and a ERM motor (M1). The motor is controlled via digital pin D8, which serves as the signal output to drive the motor. The ground (GND) of the motor is connected to the GND of the Arduino, completing the circuit. This setup enables the

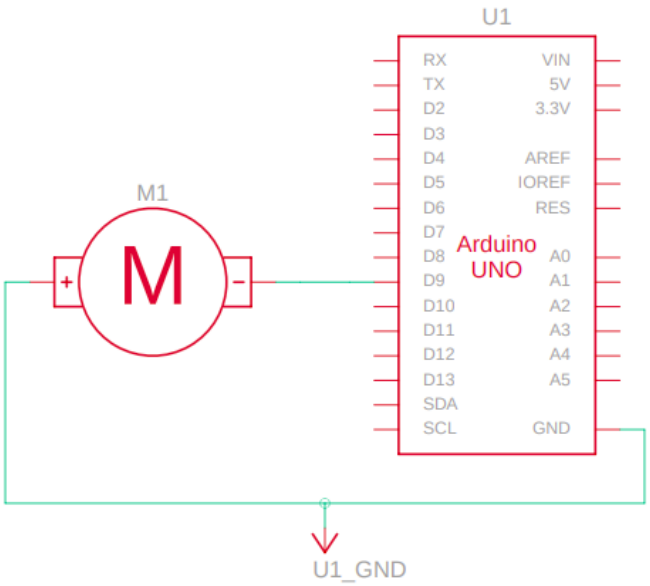

Fig. 7. Circuit Diagram

Arduino to control the motor's activation based on input signals, likely corresponding to vibration feedback in the haptic system.

10.4.3 Connections. Fig 6 showcases the real-world implementation of the system. The laptop runs the depth mapping algorithm, machine learning model, and computer vision algorithm to track finger movements and detect colors. A camera mounted on a tripod captures the interaction. The sheet serves as a reference surface where visually abled participants move their fingers, triggering different sensory feedback.

10.5 Functionalities and Baseline Implementation

10.5.1 Components/Technology Used: **Software**

- OpenCV (cv2)
- MediaPipe Hands
- NumPy,Pandas,Scikit-Learn (KNN Classifier)
- SoundDevice (sd.play)

Hardware

- Arduino UNO
- Bluetooth module
- ERM Motor
- Camera
- Breadboard

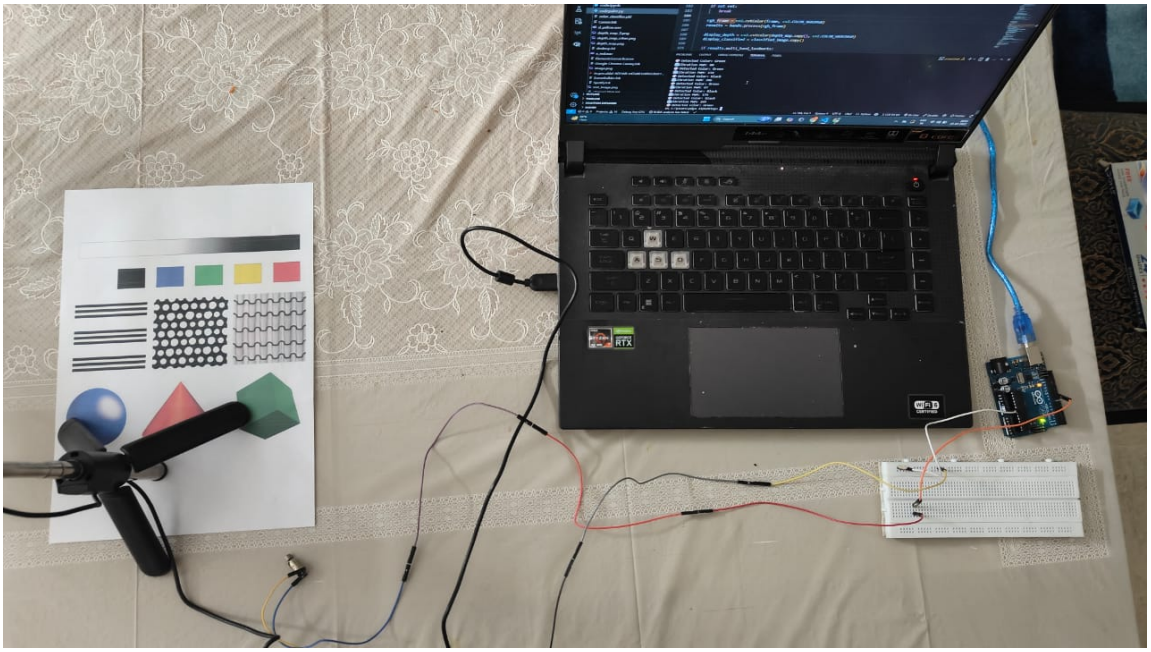

Fig. 8. Component connections

- Jumper wires

10.5.2 Haptic Feedback: Depth to Vibration. **Finger Tracking** OpenCV (cv2) is used along with MediaPipe Hands to detect and track the position of fingers in real-time.The HandTracker extracts landmark coordinates (x, y) of key points, particularly the index fingertip (landmark ID: 8). Once the hand is detected, the (x, y) position of the index fingertip is extracted from the image frame. These coordinates are normalized within the frame dimensions to ensure they fit within the required range for mapping.

Depth Mapping Algorithm Using the Depth Mapping algorithm, we generate a depth-map image where pixel intensity represents object depth. Instead of implementing a custom depth estimation model, we utilize Sculpt OK, an online depth estimation tool, as it provides a pre-trained, high-performing model. This ensures better accuracy and efficiency compared to a custom-built solution.The extracted depth data is then processed to map depth values into a scaled range suitable for haptic feedback generation.

Vibration Motor Control The processed depth data is transmitted to an Arduino microcontroller, which controls a ERM vibration motor to provide tactile feedback. The intensity of vibration is inversely proportional to the depth value—closer objects trigger stronger vibrations, while distant objects result in weaker vibrations.

The depth-map image is analyzed to determine the depth value at a specific region (e.g., the center or a tracked point). The motor vibrates with varying intensity based on that point's depth, allowing users to perceive depth through touch.

10.5.3 Audiotory Feedback: Colour to Sounds. **Training Machine Learning Algorithm** The machine learning model used for color classification is a Scikit-Learn K-Nearest Neighbors (KNN) classifier, trained on a dataset containing RGB values of six distinct colors: Red, Blue, Green, Yellow, Black, and White. Each pixel's RGB triplet served as a feature,

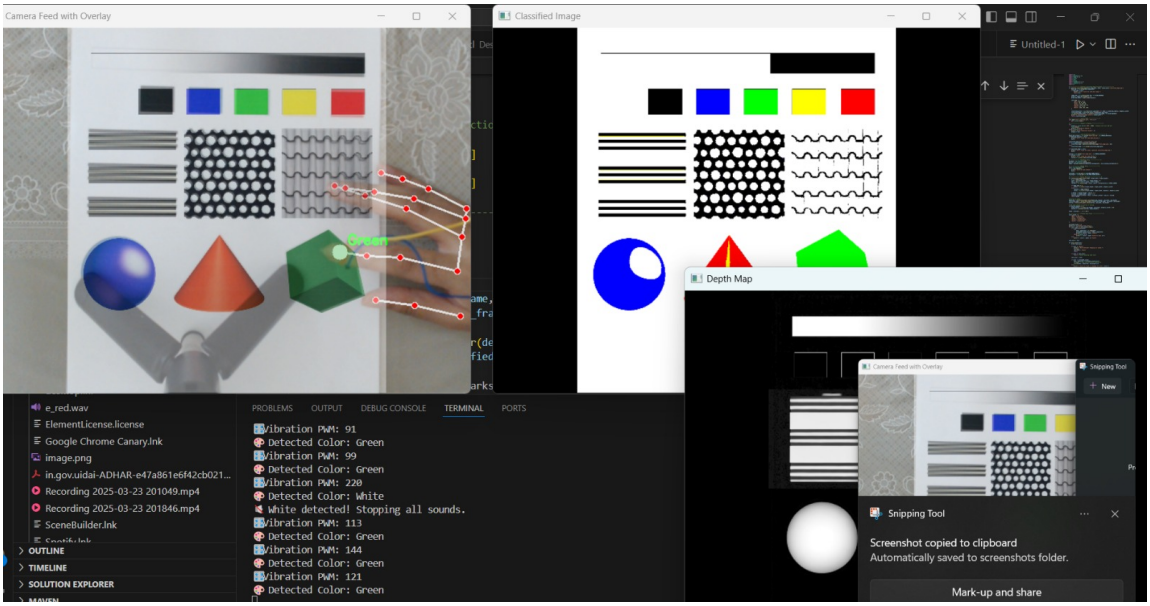

Fig. 9. Working Prototype

while the corresponding color label was the target. Then, the trained model was saved as a pickle file. This was done to ensure that complex colours can be classified into our broader colour base.

Identification of colours and Auditory mapping Once a frame is captured, it is processed using the trained KNN model, which predicts the color of each pixel. The output is then mapped to predefined audio cues that correspond to the detected colors. Each color is assigned a unique frequency pattern as described below :

- White : No Sound
- Red : E note - 659.26 Hz
- Yellow : D note - 587.33 Hz
- Green : C note - 523.25 Hz
- Blue : B note - 493.88 Hz
- Black : A note - 440 Hz

Real-Time Audio Feedback Implementation To provide continuous real-time auditory feedback, the system processes live video frames. Wherever our fingers trace the image, pixel colour values are translated to a colour label. The detected colours are converted into corresponding sound waves, which are played using Python's SoundDevice library.

10.6 Preliminary Results

Utilizing the aforementioned techniques and components, we have successfully developed a functional system. A demonstration of the system in action can be viewed in the following video: LINK (Also see Figure 7)

10.7 Evaluation Metrics

(1) Virtual Texture Perception: By leveraging haptic feedback with varying vibration intensities, the system enables users to perceive texture variations in an image without physically touching it. Different depth levels correspond to different vibration strengths, allowing users to sense surface details, edges, and contours, effectively translating visual texture into a tactile experience.

(2) Responsiveness : The system ensures seamless real-time interaction—fingers are tracked instantly, the vibration motor responds immediately to depth changes, and sound feedback is generated without delay, creating a dynamic and fluid user experience.

(3) F1 Score and Accuracy : The KNN model achieves an impressive F1 score of 0.9715, indicating a strong balance between precision and recall. We also have an accuracy of 0.9720. To further enhance performance, we can explore hyperparameter tuning.This could help refine the model's decision boundaries and improve its classification accuracy.

(4) Learnability : To improve learnability, we employ a structured mapping of sound frequencies that follows a descending tonal progression—moving from warmer to cooler tones. This gradual descent aligns with natural perceptual tendencies, making it easier for users to intuitively associate colors with their respective sounds.

(5) LLM Integration : In the future, we plan to integrate large language models (LLMs) to enhance the overall user experience and accuracy. LLMs can provide contextual assistance by dynamically adjusting sound mappings based on user preferences, offering natural language feedback, and improving object recognition through multimodal learning. This integration will make the system more adaptive, intuitive, and personalized for users.