Deep Drawing: Spectra and Sound-Source Localization

Julie Zhu
University of Michigan
zhujulie@umich.edu

Erfun Ackley
University of Michigan
ackleye@umich.edu

Zhiyu Zhang
University of Michigan
zhiyuzha@umich.edu

John Granzow
University of Michigan
jgranzow@umich.edu

ABSTRACT

This paper introduces Deep Drawing, an intermedia AI co-performer that creates a real-time artistic dialogue with human artists on a shared web-based canvas. Our system employs four contact microphones attached to a drawing surface to capture the subtle sounds of pen strokes. Upon predicting the path of the pen through custom machine learning surface sound source localization, the system overlays its predictions onto a live video feed of the human drawing. We contribute our findings on audio data pre-processing techniques, such as normalization and high-cut. We also discuss the potential differences between spectra and spectrograms for computational efficiency and prediction accuracy on this novel dataset for surface sound source localization and new interface for human-computer artistic interaction.

1. INTRODUCTION

Real-time audiovisual systems that respond to human gestures have had a long history in the performing arts. Recent advances in deep learning have further enabled new forms of expressive feedback loops between humans and machines. [1] In this paper, we focus on a drawing-based co-creative system that captures the acoustic properties of drawing on wood panels, and then interprets these signals to generate predictive strokes.

Several prior projects have shaped our perspectives on this research. From the fourth movement of Mark Applebaum's *Straitjacket* [2] where four performers draw in synchronous rhythm but their visual results are vastly different, to the *Deckle* drawing interfaces that demonstrate various ways to convert drawing into musical sounds. [3] In the domain of generative drawing, Google's SketchRNN notably models vector-based sketches and learns to produce new drawings from partial user inputs. [4] Meanwhile, large-scale datasets such as MNIST's handwritten numbers have provided a solid foundation for understanding the creative potential of working with handwriting as visual data. [5]

Our work also builds on sound source localization (SSL), which has historically involved signal processing techniques for speech recognition and robotics applications. [6] In performative contexts, these methods can be adapted to turn raw acoustic data of pen-on-paper into streams of coordinate predictions. A common method for SSL involves time difference of arrival (TDOA), but in practical applications of surface SSL on wood, the speed of sound in wood negates the usefulness of TDOA and instead, received signal strength, or the intensity of the acoustic signal as compared between microphones is the method we have employed as it also does not require time synchronization. [7]

2. METHODS

Our system design of Deep Drawing consists of four primary components: (1) a real-time multichannel audio capture and preprocessing pipeline, (2) a residual neural network for sound source localization, (3) a Kalman filter for prediction refinement, and (4) a Web interface that overlays the network's predictions onto a live video feed of the human drawing. Figure 1 illustrates the complete system design.

2.1 Data Processing

To process the audio signals, we extract both *spectrogram* and *spectral representations* from a four-channel audio recording sampled at 48 kHz. Though spectrograms are most commonly used for audio data to take advantage of computer vision models for images, we wanted to see if the more lightweight spectrum snapshots would be sufficient for our purposes. The audio data is synchronized with video recordings at a frame rate of 30 frames per second (FPS). Since the audio sampling rate is significantly higher than the frame rate, each video frame is analyzed with fine-grained temporal resolution.

2.1.1 Spectrogram Representation

For spectrogram-based processing, we compute spectrograms using the Librosa library. We apply a Short-Time Fourier Transform (STFT) with an Fast Fourier Transform (FFT) window size of 2048, a hop length of 512, and a window size of 1600. These spectrograms capture time-frequency representations of the audio signals and are generated at each video frame to preserve temporal and spectral features.

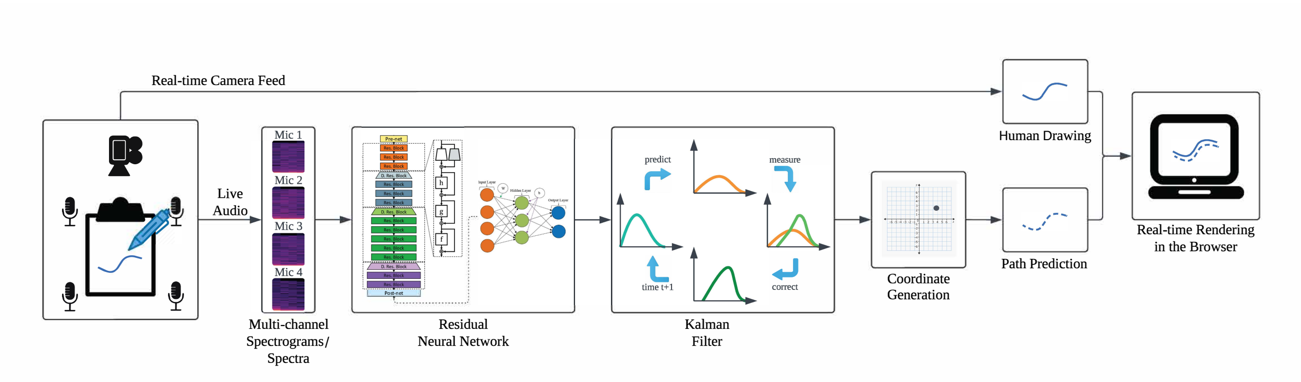

Figure 1. System design diagram of Deep Drawing.

2.1.2 Spectral Representation

For spectral analysis, we apply STFT-based spectral extraction optimized for real-time performance. To capture temporal dynamics within each frame, we divide every video frame into three sub-segments, from which we extract magnitude spectra using a 1024-point FFT.

We evaluate two configurations: (1) a full-spectrum approach, which retains all frequency bins, and (2) a high-cut version, where only low-frequency bins up to 6400 Hz are retained to focus on the most relevant spectral components for localization. This allows us to compare the effect of spectral bandwidth on model performance.

2.1.3 Silence Filtering

To exclude non-sound frames, we apply a loudness threshold of root mean square (RMS) greater than 0.005. Frames where the mean RMS amplitude across all channels falls below this threshold are marked as silent and discarded.

2.1.4 Motion Tracking and Ground Truth Extraction

The ground truth motion tracking data is obtained by detecting a pre-defined color marker on the pen tip in the video frames. The tracking algorithm operates in Hue, Saturation, and Value (HSV) color space, applying thresholding to detect lighter and darker variations of the reference color. The centroid of the largest detected contour provides the (X, Y) position of the pen.

The video frames have a resolution of 936 $\times$ 936 pixels. To ensure consistency in the training process, the extracted coordinates are normalized to a fixed range of $[0, 1]$ during data preparation. This normalization helps improve convergence during training and ensures that the model learns a scale-invariant representation of spatial positioning. The synchronized and normalized coordinates, along with their corresponding detected sound frames, are stored in a CSV file for supervised learning.

2.1.5 Dataset Size

For a dataset spanning 64.95 minutes, we extract a total of 350,730 data points from 116,910 video frames.

2.2 Neural Network Architecture

We adapt and fine-tune a pre-trained residual neural network (ResNet) to localize the artist's pen based on multi-channel audio representations. Our approach involves training both ResNet34 and ResNet50 models on two distinct input modalities: spectrograms and spectral features. While ResNet34 provides a computationally lighter alternative suitable for real-time applications, ResNet50 offers deeper feature extraction, which may enhance localization accuracy at the cost of increased complexity.

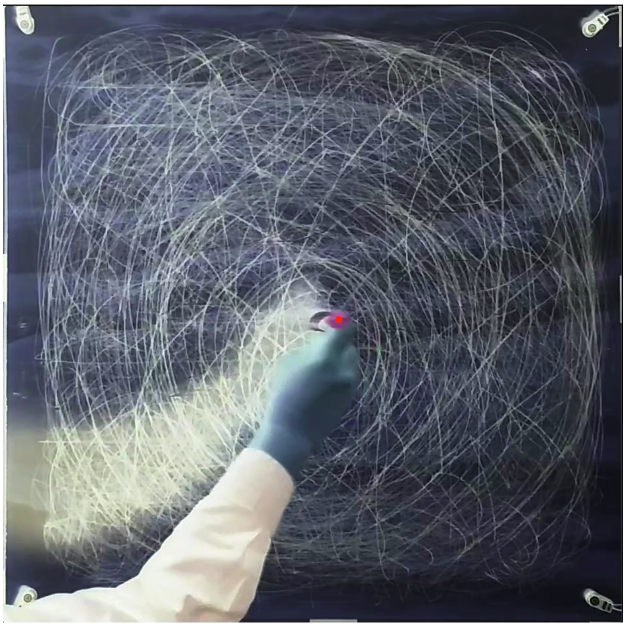

Figure 2. Visualization of the color tracking process. The red dot represents the detected pen tip position based on HSV color thresholding.

In both cases, we modify the first convolutional layer to accommodate our specific input structure. For spectrogram-based training, the input consists of spectrograms extracted from four contact microphones, necessitating an update to accept four input channels. For spectral-based training, where we extract three spectral sub-segments per frame and concatenate them into a single representation, the first convolutional layer is adjusted to accept a single-channel input.

To preserve the spatial relationships within the input representations, we retain the standard ResNet kernel configuration, using 7×7 filters with a stride of 2 and padding of 3. To prevent overfitting, we incorporate dropout regularization with $p = 0.3$ before the final fully connected layer. The network's final layer is replaced with a regression head that outputs the (x, y) coordinates of the pen on the drawing surface.

By evaluating both ResNet34 and ResNet50 across spectrogram and spectral inputs, we aim to determine the optimal trade-off between model complexity, accuracy, and real-time performance for interactive drawing applications.

2.3 Kalman Filter

We also integrate a Kalman filter that refines the predictions of the residual neural network. We model drawing motion as a constant velocity system and implement the filter's state-space model to track both position and velocity of the pen. The Kalman filter hence effectively smooths the predicted path even when individual predictions exhibit uncertainty to preserve the artistic intent of rapid directional changes. Subsequently, the system streams the Kalman filter's outputs to the Web interface through WebSocket.

2.4 Web Interface

We incorporate a React-based canvas application to present Deep Drawing as an intermedia AI co-performer to audiences. The system employs WebSocket communication to receive coordinate streams from the Kalman filter and renders them using the Perfect Freehand library to achieve pressure-sensitive stroke aesthetics redolent of stroke variation in caligraphy. Finally, we overlay the Web interface with a live video feed of the human drawing captured by a Web camera.

2.5 Implementation

We implement Deep Drawing's neural network in PyTorch and conducted all experiments on a machine equipped with an NVIDIA RTX 4090 GPU (16 GB VRAM), an Intel Core Ultra 9 185H CPU (16 cores, 22 threads), and 30 GB of RAM, running Linux with CUDA 12.4. Our optimization strategy introduces a layer-specific learning rate scheme. Specifically, we used the AdamW optimizer with a weight decay of $1e - 2$. The first convolutional layer, the batch normalization layer, and the final fully connected layer were trained with a base learning rate of $1e - 3$. Intermediate ResNet layers were trained with a lower learning rate of $1e - 4$ to better preserve pre-trained features. We adopted OneCycleLR with cosine annealing to schedule learning rates. We trained all models with a batch size of 64 for up to 200 epochs to minimize the mean-absolute error (MAE) on normalized pen-tip coordinates. Because our surface is 60×60 centimeters, the average error in centimeters is $\mathcal{L} \times 60$. We also implemented early stopping to terminate training if the validation loss plateaus for 10 consecutive epochs.

We configured Deep Drawing's Kalman filter with a 4-dimensional state space model that tracks both position (x, y) and velocity (dx, dy). Our state transition matrix modeled constant velocity motion, and our measurement matrix extracts only position components from predictions. We empirically set measurement noise (R) to $100I$, process noise (Q) to $0.1I$, and initial state covariance (P) to $1000I$. We incorporated a warm-up period of 10 predictions with relaxed filtering, followed by a distance threshold of 13.02 cm to reject outliers. For temporal consistency, we maintained a sliding window of size 5 and apply linearly increasing weights from 0.2 to 1.0 for more recent predictions.

3. EXPERIMENTS

3.1 Dataset

Four-channel WAV files were collected via four AKG 411 PP condenser contact microphones on 6mm thick 600 mm square plywood boards routed through a Clarett+ audio interface. Simultaneous pen positions were collected via overhead video routed through OBS with ground-truth minimum and maximums. The pen used was the BIC Cristal, with a sticker at the end for color-tracking, and another sticker 50 mm down for angle correction.

3.2 Effect of High-Cut on Spectral Representations

To explore the impact of filtering high-frequency components, we implemented a high-cut filter at 6400 Hz on the spectra. This approach could reduce computational overhead by focusing on low-frequency components, which have higher energy. However, we also observe that high-frequency components, though susceptible to noise and lower energy, are typically more significant for localization tasks due to high differentiation. Table 1 presents the performance comparison of models trained with and without high-cut filtering.

The inclusion of the high-cut filter resulted in a slight increase in the final loss, particularly for ResNet34, as expected. However, it offers several advantages for real-time applications. High-frequency components require finer temporal resolution and higher computational resources during both feature extraction and model inference. By limiting the frequency range, the system processes fewer data points, making it more efficient for real-time applications.

Although high-cut filtering may exclude some subtle high-frequency information, the retained low-frequency features could remain adequate for less noisy data. However, accurately localizing the artist's pen necessitates the inclusion of high-frequencies, and our higher final test losses confirms this sanity check.

Spectra + Hi-Cut	Architecture	Final Loss	Best Epoch
Y	ResNet50	0.0796	92
Y	ResNet34	0.0850	77
N	ResNet50	0.0733	99
N	ResNet34	0.0748	90

Table 1. Performance comparison of spectra with and without high-cut, evaluated on different architectures.

3.3 Comparison Across Data Types and Normalization Strategies

Table 2 shows the performance of models across spectra and spectrogram data types, with and without normalization, using ResNet34 and ResNet50 architectures. Models trained without normalization consistently achieve lower final loss, such as the ResNet50 model trained on unnormalized spectrograms with a loss of 0.0580 compared to

0.0711 with normalization. This suggests that normalization may reduce the model's ability to capture essential features for localization in this task.

Spectrogram-based models generally outperform spectra-based ones, as spectrograms better capture temporal and spatial characteristics. ResNet50 also outperforms ResNet34, with lower final loss values across most configurations. Overall, the combination of unnormalized spectrograms and the deeper ResNet50 architecture proves to be the most accurate. However, when considering live sound localization applications, spectra may be more desirable due to their potentially lower computational requirements, making them better suited for real-time processing.

Data Type	Norm	Architecture	Test Loss	Loss in cm	Best Epoch
Spectra	Y	ResNet50	0.0807	4.842	64
Spectra	Y	ResNet34	0.0944	5.664	66
Spectra	N	ResNet50	0.0733	4.398	99
Spectra	N	ResNet34	0.0748	4.488	90
Spectrograms	Y	ResNet50	0.0711	4.266	20
Spectrograms	Y	ResNet34	0.0697	4.182	18
Spectrograms	N	ResNet50	0.0580	3.480	106
Spectrograms	N	ResNet34	0.0594	3.564	112

Table 2. Comparison of performance across data types, normalization strategies, and architectures.

4. CONCLUSIONS

Our results revealed several key insights about the nature of surface sound source localization (SSL) in creative contexts. While human auditory perception may not be able to distinguish high-frequency spectral information, this information is indispensable for machine learning-based SSL. Notably, our experiments showed that normalization, which typically improve neural network training, did not improve model performance in our specific application. This suggests that the raw intensity relationships between microphone signals contain important relative spatial information that normalization may have obscured or skewed.

These technical insights inform not just the implementation of *Deep Drawing*, but also our understanding of human-AI co-creation through sound. The machine's reliance on typically imperceptible acoustic features highlights an interesting parallel to how human artists develop an intuitive understanding of their craft through extended practice. As we continue to refine our system for real-time performance, these findings will guide our optimization of the balance between prediction accuracy and latency. Our future work will evaluate how these design decisions influence the quality of interaction between human artists and the AI co-performer in live drawing sessions.

5. DISCUSSION AND FUTURE WORK

Surface sound source localization has seen notable applications in fields such as seismology, particularly for earthquake detection. In these contexts, arrays of sensors—seismometers—are deployed to detect and locate seismic events based on wave propagation across the Earth's crust. [8] Although the underlying principles of wave-based localization parallel those in our setup with contact microphones on a wooden surface, there has been relatively limited exploration of surface SSL in creative domains. Leveraging recent advances in deep learning for real-time signal processing, our system brings SSL beyond its traditional use cases. This shift from geophysical monitoring to intermedia human-AI co-creation raises intriguing questions about how seemingly noisy signals can become rich conduits for creative expression.

More data is needed for a robust model, and in a few directions. The type of drawing utensil can be varied, so that the model can accommodate any spectral signature of a writing gesture. We have begun this data collection to include pencil, charcoal, and graphite stick.

In addition, the human making the drawing should be varied as much as possible. Each person has a unique way of drawing, and while the model may be fine-tuned to an individual, the hope is that it can be applicable to any performer.

Currently, the model architecture utilizes the well-known convolutional neural network (CNN) ResNet. While we expect CNNs to remain the fastest solution for real-time surface SSL, we plan to investigate the equally promising alternatives in our future work: FCNNs, LSTMs, and transformers applied to 1D spectral data [9].

For this type of model, a faithful reproduction of someone's signature from just its surface aural signal, say, is worth the wait.

Certainly, the purpose of this model is not to steal people's signatures, but rather to understand how to process a very noisy signal. Our current application is performance-based, with the goal of understanding the human writing and drawing gesture through trying to teach a machine to do the same and watching the machine generally fail.

Acknowledgments

Deep Drawing is supported by the Performing Arts Technology Department, the ADVANCE program, and ArtsEngine at the University of Michigan.

6. REFERENCES

[1] E. R. Miranda, *Handbook of artificial intelligence for music*. Springer, 2021.

[2] M. Applebaum, "Straitjacket," 2019, score, retrieved from https://web.stanford.edu/ applemk/portfolio-works-straitjacket.html.

[3] H. Choi, J. Granzow, and J. Sadler, "The Deckle Project: A Sketch of Three Sensors." in *NIME*, 2012.

[4] D. Ha and D. Eck, "A Neural Representation of Sketch Drawings," 2017. [Online]. Available: https://arxiv.org/abs/1704.03477

[5] P. Grother and K. Hanaoka, "NIST special database 19 handprinted forms and characters 2nd edition," *National Institute of Standards and Technology, Tech. Rep*, vol. 5, 2016.

[6] G. Jekateryńczuk and Z. Piotrowski, "A Survey of Sound Source Localization and Detection Methods and Their Applications," *Sensors*, vol. 24, no. 1, p. 68, 2023.

[7] Y. Li, S.-S. Yu, L. Dai, T.-F. Luo, and M. Li, “Acoustic emission signal source localization on plywood surface with cross-correlation method,” *Journal of Wood Science*, vol. 64, no. 1, pp. 78–84, 2018.

[8] R.-S. Jia, Y. Gong, Y.-J. Peng, H.-M. Sun, X.-L. Zhang, and X.-M. Lu, “Time difference of arrival estimation of microseismic signals based on alpha-stable distribution,” *Nonlinear Processes in Geophysics*, vol. 25, no. 2, pp. 375–386, 2018.

[9] Y. Sun, S. Brockhauser, and P. Hegedűs, “Comparing End-to-End Machine Learning Methods for Spectra Classification,” *Applied Sciences*, vol. 11, no. 23, p. 11520, 2021.