Library Progress International **Print version ISSN 0970 1052**
Vol.44 No. 3, Jul-Dec 2024: P. 11997-12011 **Online version ISSN 2320 317X**

Original Article **Available online at www.bpasjournals.com**

Exploring Audio-Visual Correlation for Real-Time Texture Analysis

[1]Shyam Maheshwari*, [2]Dr. Hemant Makwana, [3]Dr.Devesh Kumar Lal

Author's Affiliation:

[1]Institute of Engg. and Tech.,Devi Ahilya University, Indore, 452001, Madhya Pradesh, India
[2]Institute of Engg. and Tech.,Devi Ahilya University, Indore, 452001, Madhya Pradesh, India
[3] Madhav Institute of Technology and Science., Gwalior, 457001, Madhya Pradesh, India

This study presents a novel exploration of the correlation between surface images and audio generated by linear movement over surfaces using three key similarity metrics: Euclidean Distance (ED), Cosine Similarity, and Pearson Correlation. A key contribution of this research is the application of these metrics to the LMT TUM Texture dataset, revealing new insights into their comparative effectiveness for multisensory fusion. The results demonstrate that Cosine Similarity and Pearson Correlation maintain high stability across varying surface textures, making them ideal for real-time applications in augmented reality (AR), human-robot interaction (HRI), and industrial monitoring. In contrast, Euclidean Distance exhibits greater sensitivity to texture changes, highlighting its utility for detecting subtle variations in surface properties, especially for fault detection and industrial applications. The study also identifies how increasing the Sigma value enhances similarity, as Euclidean Distance decreases while Cosine Similarity and Pearson Correlation approach near-perfect correlation. Although the analysis is insightful, the research is limited by the narrow range of surface textures and the absence of real-time implementation. Future research will expand the dataset and integrate machine learning techniques to enhance real-time performance. This work advances the field by offering a robust framework for understanding and optimizing multisensory fusion systems across practical applications.

KEYWORDS
Surface Texture Analysis, Multimodal Data Fusion, Similarity Matrices, Audio-Visual Correlation, Human Computer Interaction

1. Introduction

The correlation between surface images and audio generated by movement over surfaces represents a growing area of research within the broader field of information fusion, which integrates data from multiple sensory modalities for improved decision-making and system accuracy. This field is particularly relevant in applications where multisensory feedback is critical, such as robotics, augmented reality (AR), virtual reality (VR), and material science. By understanding the relationship between auditory and visual data, researchers can enhance the realism of simulations, improve robotic tactile sensing, and develop more accurate systems for navigation, classification, and fault detection.

Recent advancements have focused on utilizing image similarity measures to analyze non-traditional image types, such as Melspectrograms derived from audio signals. These spectrograms provide a visual representation of audio data, effectively transforming time-frequency components into images that can be compared using well-established image metrics. This novel approach has shown great potential in various tasks, including industrial sound classification, music genre identification, and fault detection in mechanical systems. By

converting audio data into a format amenable to visual analysis, researchers can leverage powerful image processing techniques to un- cover patterns and relationships that might otherwise remain hidden. In these contexts, metrics like Euclidean Distance (ED), Pearson Correlation Coefficient (PCC), and Cosine Similarity have been employed to assess the descriptive strength of images and their ability to differentiate between distinct sound classes, thus enhancing our understanding of the underlying audio phenomena.

The use of Mel spectrograms and advanced image comparison techniques has been particularly successful in industrial applications, where the sound generated by movement over surfaces can provide valuable insights into surface texture and material properties. This research explores the intricate relationship between surface images and their corresponding audio signals, aiming to improve real-time classification systems through multisensory fusion. By integrating visual and auditory data, we not only gain a better understanding of surface textures but also enhance the overall system's performance. This fusion of modalities is not only beneficial for recognizing surface characteristics but also plays a crucial role in enhancing the accuracy and reliability of human-robot interaction (HRI) and autonomous navigation systems. Such advancements can lead to more intuitive interfaces and smarter robotic systems capable of operating effectively in complex environments, ultimately paving the way for greater automation and efficiency in various industries.

This paper reviews the key methods and metrics used to correlate surface images with audio, with a particular focus on ED, PCC, and Cosine Similarity. These techniques are evaluated based on their performance in various applications, including real-time monitoring, fault detection, and classification tasks. By providing a com- prehensive overview of the state of the art in this field, this study aims to identify gaps in current methodologies and suggest future directions for research that can further ad- vance multisensory fusion systems, contributing to the broader goals of information fusion.

2. LITERATURE REVIEW

The study of correlating surface images with audio generated by movement over surfaces has emerged as a multidisciplinary field encompassing image processing, audio signal analysis, and multisensory fusion. Various similarity metrics, such as Euclidean Distance (ED), Pearson Correlation Coefficient (PCC), and Cosine Similarity, have been employed to explore the relationship between visual and auditory data. This area of research is particularly relevant in applications such as human-robot interaction (HRI), augmented reality (AR), virtual reality (VR), and other sensory substitution environments, where accurate navigation and guidance systems are crucial [1].

2.1. Euclidean Distance (ED)

Euclidean Distance (ED) is one of the simplest and most widely used measures in image processing and data analysis. ED calculates the straight-line distance between two points in a multi-dimensional space, quantifying image similarity by comparing pixel-wise differences. The lower the ED, the more similar the images are. ED has been employed in various applications, including stereo pair matching, visual tracking, and image retrieval. In stereo vision, ED helps align images by minimizing the distance between corresponding pixels [2]. This technique can also be extended to audio analysis, where ED is used to compare spectrograms generated from audio signals to assess the similarity between sounds produced by different surface textures [3].

Despite its computational simplicity, ED's reliance on magnitude makes it sensitive to outliers and less effective in high-dimensional, noisy environments. This limitation suggests the need for complementary measures like PCC and Cosine Similarity for more complex audio-visual correlations [4].

2.2. Pearson Correlation Coefficient (PCC)

PCC measures the linear correlation between two datasets, offering insights into how closely two variables are related. PCC has been used in image analysis to assess correlations between pixel intensities, making it particularly effective for tasks requiring linear relationships, such as medical imaging and remote sensing [5]. In audio analysis, PCC can quantify the correlation between sound features and their corresponding surface textures by comparing Mel-spectrograms, thereby improving the accuracy of sound-based classification in real-time systems [6].

PCC's application in fault detection systems, particularly in industrial environments, has shown its effectiveness

in classifying sounds generated by surface interactions. The measure helps capture linear relationships between sound and surface texture data, enhancing the robustness of surface texture recognition systems [7]. However, PCC may struggle with non-linear relationships and requires complementary measures to fully capture the complexity of visual and auditory data [8].

2.3. Cosine Similarity

Cosine Similarity measures the cosine of the angle between two vectors, focusing on their orientation rather than magnitude, which makes it particularly useful in high- dimensional data analysis such as image and audio spectrograms. Unlike ED, which is sensitive to scale, Cosine Similarity is invariant to magnitude differences, allowing it to capture structural similarities between datasets [9]. In audio-based systems, Co- sine Similarity is commonly employed to evaluate structural similarities between sound signatures, making it a valuable tool for comparing Mel spectrograms generated from surface interactions. This measure has been widely used in machine learning and deep learning applications, where it is applied to high-dimensional feature vectors extracted from audio data [10]. Its robustness in noisy environments has made Cosine Similarity especially useful for cross-modal studies where visual and auditory data are combined to enhance classification and recognition tasks [11].

2.4. Multisensory Fusion and Real-World Applications

The integration of image and audio data through multisensory fusion techniques has shown great promise in various applications, including autonomous vehicle navigation, HRI, and AR/VR systems. Multisensory fusion combines different data types (e.g., visual, auditory, and tactile) to provide a comprehensive understanding of the environment [12]. This approach has been instrumental in systems that classify surface textures based on the sounds generated by movement over those surfaces [13]. In industrial monitoring, combining visual and auditory data improves the accuracy of fault detection and anomaly classification systems. For instance, systems that monitor mechanical sounds and align them with surface textures offer enhanced detection capabilities compared to systems that rely on audio data alone [14]. Additionally, advancements in deep learning techniques have enabled more accurate real-time data fusion from multiple sources, leading to better performance in recognizing and classifying surface properties [15].

2.5. Advanced Methods and Applications

The use of machine learning techniques, such as convolutional neural networks (CNNs), has led to innovations in audio recognition based on mel-spectrograms [19]. This has opened the door for further exploration in fault detection and industrial monitoring, where sound can be used to detect surface wear and other anomalies [20].In recent studies, hyperspectral image classification using sharpened Cosine Similarity operations has shown promise in improving the accuracy of object detection [21]. Similar approaches can be applied to correlate surface textures and movement-generated sounds, as they rely on high-dimensional data relationships between audio spectrograms and image features. The fusion of audio and image data can also benefit from innovative approaches like the use of tiny neural networks to approximate spectrogram features in real-time applications [22]. For example, in-home appliance classification, neural networks have shown their effectiveness in handling multiple data types simultaneously. This approach can be adapted to more complex audio-visual fusion scenarios. Additionally, methodologies for bearing vibration investigation based on spectrogram image comparison [23] and remaining useful life (RUL) prediction [24] have demonstrated the effectiveness of using audio-visual data fusion in predictive maintenance and industrial applications. These studies highlight the increasing relevance of multi-sensory data fusion in industrial systems, particularly for real-time anomaly detection and equipment monitoring [27-32].

Future directions for research should also explore new methods for surface visualization using audio images. Techniques for detecting tool wear through sound analysis [25], as well as similarity-based correlation algorithms for object detection [20], represent cutting-edge applications that bridge the gap between image and audio data fusion.

2.6. Research Gaps

Despite advancements in image and audio correlation, several challenges persist. One significant issue is the lack of universally accepted standards for measuring image similarity, with the choice of metrics varying across

applications [18]. Furthermore, audio-based classification and recognition methods remain less mature than image- based techniques, particularly in areas such as fault detection and industrial monitoring [16]. Real-time multisensory fusion remains a key area for further research, with an emphasis on improving accuracy and computational efficiency in practical applications. Future work should focus on integrating machine learning and deep learning models with existing similarity metrics, such as ED, PCC, and Cosine Similarity, to enhance real-time multimodal fusion systems. Advances in these areas will be critical for im- proving the robustness and reliability of systems in fields ranging from autonomous navigation to immersive VR environments [17].

3. DATASET

The TUM LMT Haptic Dataset 108 is a comprehensive resource designed to advance research in haptic feedback and surface characterization. It includes 108 care- fully selected objects representing a wide variety of surface textures and material properties, organized into nine distinct groups for detailed analysis. Published by Sress et al. at Technische Universita¨tMu¨nchen (TUM), the dataset comprises two primary data sets: one acquired under controlled conditions with constant parameters except for force and scanning velocity, and another captured through ten free-hand recordings (five linear and five circular movements). The dataset features images, three-axis acceleration signals, sound signals, and friction signals, but this research specifically focuses on images and audio. The texture images are categorized into groups such as mesh, stones, glossy surfaces, wood, rubbers, fibers, foams, foils, and textiles. The file structure facilitates ease of access, with high-resolution images (320 × 480 pixels) and audio recordings for each object organized systematically.

In this study, the dataset is leveraged to explore the correlation between visual and auditory data from surface textures. The 20 images for each object are converted into audio signals using image sonification techniques, resulting in Mel spectrograms for both images and audio recordings. By employing quantitative measures like Euclidean distance, cosine similarity, and Pearson correlation, the research aims to uncover significant relationships between the surface images and the audio generated from linear movements across these surfaces. Overall, the TUM LMT Haptic Dataset 108 provides a solid foundation for analyzing the interplay between visual and auditory characteristics of surfaces.

4. SCOPE AND METHODOLOGY

As per our motivation to identify texture properties through sound we need to classify audio files based on texture classification. That will be helpful for us to recognize the roughness and smoothness level of the product. to identify that we used two methods- ologies. The first one is for using audio file properties itself we use sound classification for classify that files. The second one is to use a spectrogram image of the audio files for image-based classification. By using both methodologies we want to find fast and efficient way to classify texture through audio.

4.1. Image to Spectrogram Conversion

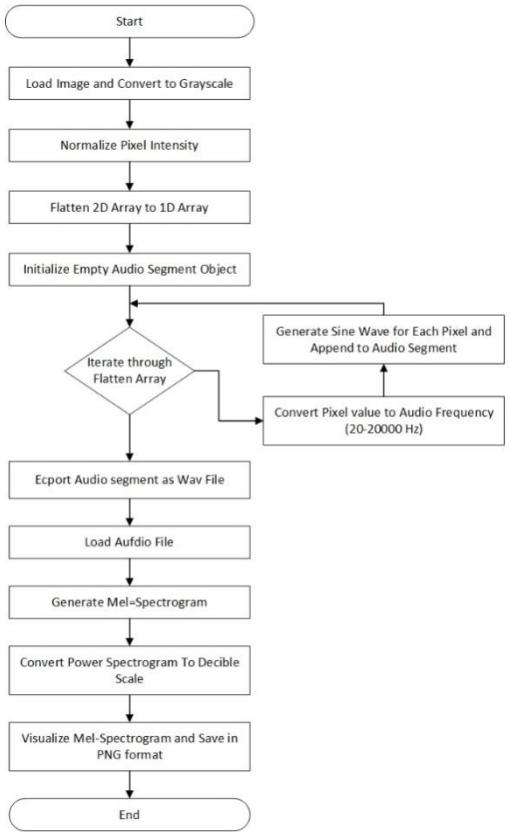

Figure 1: Methodology of image-based classification

The process begins with loading an image and converting it to grayscale using thePython Imaging Library (PIL). This reduces the image to a single channel, with pixel intensity values ranging from 0 (black) to 255 (white). Next, these intensity values are normalized to a range between 0 and 1, which standardizes the data and facilitates consistent conversion of pixel values to audio frequencies. The normalized two-dimensional array of pixel values is then flattened into a one- dimensional array. Flattening the array prepares the data for sequential audio generation, making it easier to iterate through each pixel value. An empty Audio-Segment object is initialized, which will hold the concatenated audio signals generated from the pixel values.

The script then iterates through the flattened array of pixel values. This iteration allows for the conversion of each pixel value into a corresponding audio frequency. Each pixel value is converted to an audio frequency within the range of 20 Hz to 20,000 Hz, based on a base frequency of 440 Hz (A4 note) and a pitch shift applied according to the pixel intensity. This step translates visual information into auditory signals, which is the core of the sonification process.

A sine wave of the calculated frequency is generated for each pixel.Sine and ap- pended to the audio segment. The duration of each sine wave corresponds to the duration per pixel parameter. Generating and appending sine waves creates an audio representation of the image, with each pixel's frequency contributing to the overall sound. The concatenated audio segment, representing the entire image, is then exported as a WAV file. Exporting the audio file enables further analysis and comparison with movement-derived audio files.

The generated audio file is first loaded, allowing the audio data and sample rate to be read. This step prepares the audio file for conversion into a Mel spectrogram. The Mel spectrogram is then created, computing a

spectrogram where the frequency axis is transformed to the Mel scale. This process utilizes 128 Mel bands and a maximum frequency of 8000 Hz, providing a detailed time-frequency representation of the audio signal.

Next, the power spectrogram is converted to a decibel (dB) scale. This conversion uses a logarithmic scale to enhance the visualization of amplitude variations, making the data easier to interpret. The Mel spectrogram is then visualized, showing the spectrogram with time on the x-axis and Mel frequencies on the y-axis. A color bar is added to represent the amplitude in dB, and the plot is saved as a PNG image file. This visualization and saving of the Mel spectrogram enable further analysis and comparison with other spectrograms, facilitating the study of correlations between visual and auditory properties.

4.2. Audio to Spectrogram Conversion:

For Audio to Spectrogram Conversion, the process can be outlined as follows:

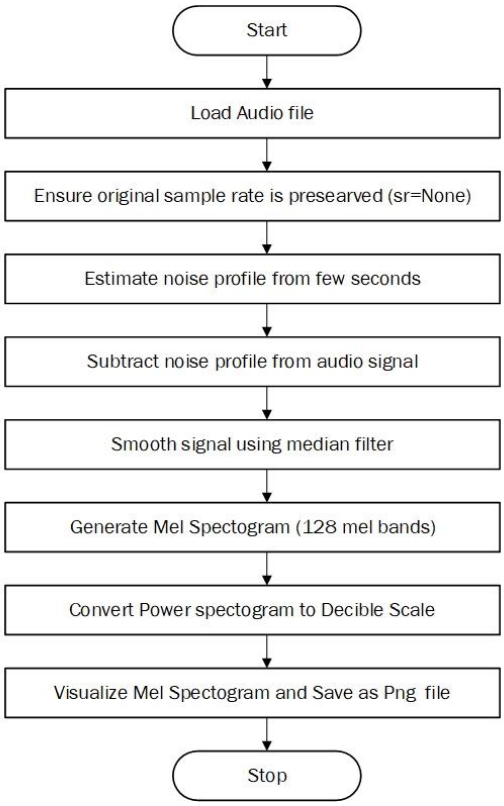

Figure 2: Methodology of Audio-based classification

The process begins with initializing the workflow, followed by loading all audio files from the specified directory. During this stage, the original sample rate of the audio files is preserved to maintain data integrity. The next step involves noise estimation and removal, where a noise profile is estimated using the initial few seconds of the first audio file, which is presumed to contain only noise. This noise profile is then subtracted from the audio signal to reduce unwanted noise. To further refine the audio quality, the noise-reduced signal is smoothed using a median filter, which helps eliminate any residual noise.

Once the audio signals are cleaned, Mel spectrograms are generated from these noise-reduced files. This involves converting the audio signal into a spectrogram with 128 Mel bands and a maximum frequency of 8000 Hz. The power spectrogram is then transformed to a decibel scale, enhancing the perceptual relevance of amplitude variations. These Mel spectrograms are visualized to provide a time-frequency representation and are subsequently saved as PNG images in the output directory. The directory structure of the output is designed to

mirror that of the input, ensuring organized storage. The workflow concludes with the completion of the spectrogram generation and saving process.

4.3. Quantitative Analysis

The first step in the procedure is loading and preparing the photographs. First, every image is loaded and grey-scaled. To improve contrast, histogram equalization is then used. The image is then normalized to scale the pixel values to a range of [0, 1] after a Gaussian blur has been used to smooth the image.

Next, similarity metrics between pairs of images are computed. This involves flattening the image arrays and standardizing the data. Several metrics are calculated, including Euclidean distance, cosine similarity, Pearson correlation, SSIM index, PSNR value, and HOG similarity. Different insights about the similarity between the photos are offered by each of these metrics.

The script then processes the images by reading an input CSV file that contains details about the image files. The images are grouped based on specific criteria, and pairs of images are selected for comparison if they belong to different folders. Each pair of images is loaded and pre-processed, and the similarity metrics are computed and recorded. These results, along with computation times for each metric, are stored in a data frame.

Finally, the collected results are saved to an output Excel file. This structured process ensures that images are efficiently compared, and the similarity metrics are accurately recorded for further analysis. The overall methodology provides a comprehensive approach to image pre-processing and similarity computation, facilitating detailed comparisons within the dataset.

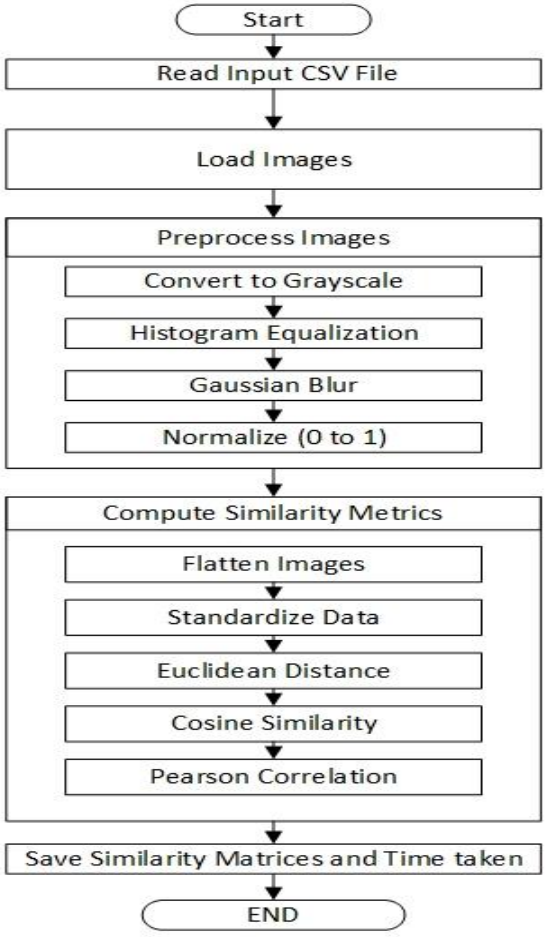

Figure 3: Methodology of similarity Matrices

The Mel spectrograms obtained from the image-derived audio and the movement- derived audio were analyzed using three robust metrics: Euclidean Distance, Cosine Similarity, and Pearson Correlation. Each metric provides unique insights into the relationship between the visual and auditory representations of the surfaces.

1. Euclidean Distance:

Description: Euclidean Distance is an estimation of the length of a straight line between two locations in the Mel spectrogram space. It quantifies the absolute difference between the corresponding elements of the spectrogram matrices. The formula for Euclidean Distance d between two vectors X and Y is:

$$D = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

Application: In our analysis, Euclidean Distance helps identify the degree of dissimilarity between the Mel spectrograms generated from images and those from audio recordings. A smaller Euclidean Distance indicates a higher degree of similarity between the two spectrograms, suggesting that the image-derived audio closely matches the movement-derived audio.

2. Cosine Similarity:

Description: The Cosine Similarity measure determines the direction of alignment between two vectors by calculating the cosine of their angle. It emphasizes the orientation of the vectors rather than their size, in contrast to Euclidean Distance. The formula for Cosine Similarity S between two vectors X and Y is:

$$S = \frac{X \cdot Y}{|X||Y|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \sqrt{\sum_{i=1}^{n} Y_i^2}}$$

Application: By calculating the Cosine Similarity between the Mel spectrograms, we measure how closely the shapes of the spectrograms align. High Cosine Similarity values indicate that the two spectrograms share similar patterns and structures, reinforcing the correlation between the image-derived and movement-derived audio.

3. Pearson Correlation:

Description: Pearson Correlation evaluates the linear correlation between two datasets, providing a measure of their co-variation. Its values vary from -1 to 1, where 0 denotes no linear relationship, -1 denotes a perfect negative linear relationship, and 1 represents a flawless positive linear relationship.

Application: In our research, Pearson Correlation quantifies the degree to which changes in the image-derived Mel spectrogram correspond to changes in the movement-derived Mel spectrogram. High Pearson Correlation values suggest a strong linear relationship, implying that variations in the visual surface characteristics are consistently reflected in the auditory properties.

Calculation of Pearson Correlation Coefficient: The formula for calculating
Pearson's r is:

$$r = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma(X_i - \overline{X})^2 \Sigma(Y_i - \overline{Y})^2}}$$

The sample points indexed with i are denoted byXi and Yi, whereas the means of the X and Y variables are represented by X and Y.

By calculating Euclidean Distance, Cosine Similarity, and Pearson Correlation, we gain a comprehensive understanding of the similarity between Mel spectrograms from different sources. These metrics collectively help us determine how well the audio characteristics derived from surface texture images align with those

generated by actual movement, providing valuable insights into the correlation between visual and auditory representations of surfaces.

4. RESULT AND DISCUSSION

By implementing this methodology, we ensure a systematic and detailed approach to pre-processing images and computing similarity metrics, which is crucial for rigor- ous image analysis in texture mapping. The saved CSV files from this methodology contain essential details such as the file name, Euclidean Distance, Cosine Similarity, and Pearson Correlation, along with the times taken to load and process each file, as well as the duration required to calculate the similarity matrices. This comprehensive data collection allows for thorough evaluation and facilitates meaningful comparisons.

Specifically, we have compared the surface images and the audio generated by moving over those surfaces, resulting in a total of 2,160 comparisons for each sigma value. For a sigma value of 10, this leads to an impressive 21,600 comparisons, from which we extracted insightful tabular data as examples. The analysis includes comparisons between image-derived audio and movement-derived audio, yielding detailed tables that illustrate the relationships identified through our methodology.

Euclidean Distance	Cosine Similarity	Pearson Correlation	Sigma Value
396.5674	0.803216	0.803216	0.5
377.0625	0.821409	0.821409	1
353.1891	0.841873	0.841873	2
335.3112	0.856505	0.856505	3
307.5745	0.878061	0.878061	5
259.0344	0.912669	0.912669	10
208.367	0.944652	0.944652	20
130.5137	0.980981	0.980981	50
95.95543	0.992991	0.992992	100
71.16455	0.99678	0.99678	200

Table 1: Similarity Measures for G1 Fine Aluminium Mesh

Figure for table 1 G1 Fine Aluminum Mesh

Surface Class	Euclidean Distance	Cosine Similarity	Pearson Correlation	Rough rank
G3 Acrylic Glass	120.0796	0.984431	0.984521	1
G5 Rubber Plate Version 2	106.2779	0.987714	0.987949	2
G1 Fine Aluminum Mesh	71.16455	0.99678	0.99678	3
G9 Table Cloth Version 1	89.81144	0.993949	0.994012	4
G4 Compressed Wood Version 2	53.76747	0.9978	0.997905	5
G6 Sheep Skin	108.9763	0.990059	0.990146	6
G8 Glitter Paper Version 1	104.8031	0.996302	0.996303	7
G2 Stone Tile Version 3	84.58768	0.995307	0.995341	8
G7 Coarse Foam	115.3624	0.986901	0.986923	9

Table 2: Correlation table between image and audio (Sigma Value=200)

Rough Rank	Explanation
1	Appears to be the smoothest surface with no visible texture, rank 1
2	Slightly textured but still relatively smooth, ranking 2
3	Has a fine texture but still fairly smooth, ranking 3
4	Woven texture with a moderate amount of roughness, ranking 4
5	Visible texture and slightly rough, ranking 5
6	Notice able texture and roughness, ranking 6
7	Rough due to the glittery surface, ranking 7
8	Rough texture with noticeable grains, ranking 8
9	Very rough texture, ranking 9

Table 3: Roughness ranking table as surface property

5. DISCUSSION

5.1. Euclidean Distance with Sigma Value

Graph Description: This graph Figure 4 shows how the Euclidean Distance between the two files changes as the Sigma Value increases.

Observations:

The Euclidean Distance decreases as the Sigma Value increases.

2. The relationship appears to be non-linear, with a steeper decline at lower Sigma Values and a more gradual decline at higher Sigma Values.

3. This suggests that as the Sigma Value increases, the Euclidean Distance between the two files decreases, indicating the files become more similar in terms of their Euclidean Distance.

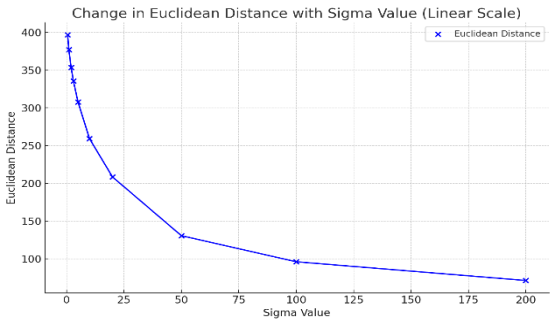

Figure 4: Euclidean Distance with Sigma Value graph

5.2. Cosine Similarity with Sigma Value

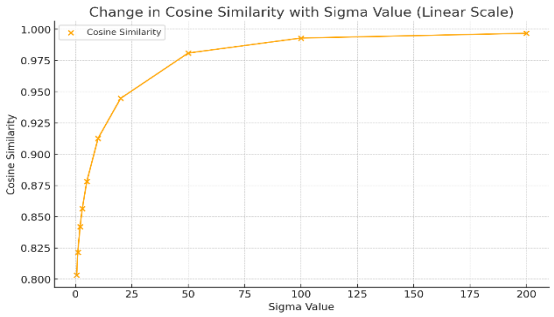

Figure 5: Cosine Similarity Change with Sigma Value graph

Graph Description: This graph figure 5 illustrates the change in Cosine Similarity be- tween the two files as the Sigma Value increases.

Observations:

1. The Cosine Similarity increases with an increase in Sigma Value.
2. The increase is quite rapid initially and then gradually approaches 1.
3. Higher Sigma Values lead to higher Cosine Similarity, indicating that the files become more similar in terms of the angle between their vector representations.

5.3. Pearson Correlation with Sigma Value

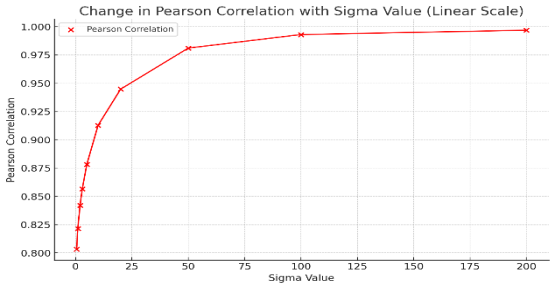

Figure 6: Pearson Correlation with Sigma Value graph

Graph Description: This graph depicts the change in Pearson Correlation be- between the two files as the Sigma Value increases.

Observations:

1. Similar to the Cosine Similarity, the Pearson Correlation also increases with an increase in Sigma Value.
2. The correlation rapidly increases initially and then starts to level off as it approaches 1
3. A higher Sigma Value corresponds to a higher Pearson Correlation, showing that the files become more similar in terms of their linear relationship.

5.4. General Insights

Similarity Trends: Both Cosine Similarity and Pearson Correlation show a clear trend of increasing similarity with higher Sigma Values, suggesting that the two metrics behave similarly in this context.

Euclidean Distance: The Euclidean Distance metric shows an inverse trend compared to the similarity metrics, with the distance decreasing as the Sigma Value increases, indicating greater similarity at higher Sigma Values

These graphs help to visualize how changes in the Sigma Value affect different measures of similarity between the files, providing a comprehensive understanding of the relationships between these metrics.

5.5. Understanding Smoothness to Roughness Ranking in Comparison to Pearson Correlation, Cosine Similarity, and Euclidean Distance.

The table provided includes the smoothness-to-roughness ranking for various sur- faces and compares them with three key metrics: Pearson Correlation, Cosine Similarity, and Euclidean Distance. Here's a detailed understanding of the relationships between these metrics and the smoothness-to-roughness ranking.

5.5.1. Detailed Analysis

1. Smoothest Surface (G3AcrylicGlass, Rank 1):

- Euclidean Distance: 120.08 (highest distance, least similarity by this metric)
- Cosine Similarity: 0.9844 (high similarity)
- Pearson Correlation: 0.9845 (high similarity)

Smoother surfaces show higher Euclidean Distance values, indicating less similarity, but high Pearson and Cosine values indicating strong linear and angular relationships.

2. Roughest Surface (G7CoarseFoam, Rank 9):

- Euclidean Distance: 115.36 (lower distance, more similarity by this metric)
- Cosine Similarity: 0.9869 (high similarity)
- Pearson Correlation: 0.9869 (high similarity)

Rougher surfaces tend to have lower Euclidean Distance values, indicating higher similarity in terms of direct distance, while Pearson and Cosine values remain high.

3. Intermediate Surfaces (Ranks 2-8):

- Euclidean Distance: Values vary but generally decrease with increasing roughness.
- Cosine Similarity and Pearson Correlation: Remain consistently high across all textures, showing strong relationships irrespective of surface roughness.

For example, G4CompressedWoodVersion2 (Rank 5) has an Euclidean Distance of 53.77, indicating high similarity, and both Cosine Similarity (0.9978) and Pearson Correlation (0.9979) are very high.

5.5.2. Insights:

Smoothness to Roughness Ranking is inversely related to Euclidean Distance: Smoother surfaces tend to have Higher Euclidean Distance values are typically found for smoother surfaces, whilst lower values are found for rougher surfaces..

Cosine Similarity and Pearson Correlation are consistently high across different textures, indicating that these metrics are less sensitive to texture changes and capture strong relationships between image and sound data regardless of surface texture.

The smoothness-to-roughness ranking provides a qualitative measure of surface texture, while the quantitative metrics (Euclidean Distance, Cosine Similarity, and Pearson Correlation) offer insights into the similarity between image and sound data. Euclidean Distance varies more with surface texture, indicating its

sensitivity to texture changes, while Cosine Similarity and Pearson Correlation remain robust, showing strong relationships across all texture

6. CONCLUSION

This study presents a thorough analysis of the correlation between surface images and audio generated by linear movement across various surfaces, using three key similarity metrics: Euclidean Distance (ED), Cosine Similarity, and Pearson Correlation. Each of these metrics provided valuable insights into how visual and auditory data can be linked, offering a robust framework for multisensory fusion systems.

Result Summary: The findings reveal that both Cosine Similarity and PearsonCorrelation consistently exhibited high values across all surface textures, indicating their robustness in maintaining strong relationships between image and sound data regardless of texture changes. These metrics were relatively insensitive to variations in surface textures, making them ideal for applications requiring stable correlations, such as in real-time data fusion for AR/VR and human-robot interaction (HRI). In contrast, Euclidean Distance showed greater sensitivity to texture changes; smoother surfaces had higher Euclidean Distance values, suggesting less similarity, while rougher surfaces indicated greater similarity. This characteristic makes Euclidean Distance particularly valuable for applications focused on detecting subtle changes in surface properties, such as industrial monitoring and fault detection systems.

Strengths: One of the main strengths of this study is its comprehensive analysis through the employment of three distinct similarity metrics. Each metric offers a unique perspective on the relationship between surface images and movement-generated audio, creating a well-rounded framework for analyzing multimodal data. Euclidean Distance effectively captures texture variations, making it particularly valuable for applications that require sensitivity to surface properties. In contrast, Cosine Similarity and Pearson Correlation provide strong, stable metrics that ensure consistent performance across a range of textures, making them highly suitable for real-time data fusion applications. Moreover, the flexibility of this methodology allows for its application in various fields, including human-robot interaction, augmented and virtual reality, and industrial monitoring. Ultimately, this study demonstrates that combining different similarity metrics can significantly enhance the accuracy and effectiveness of real-time data fusion systems that depend on correlating image and audio data, paving the way for more sophisticated multimodal analyses.

Weaknesses: Despite the valuable insights gained, the study has limitations. One key issue is the narrow range of surface textures tested; expanding the dataset to include a broader variety would enhance understanding of these metrics in diverse real-world ap- plications, particularly in fields like robotics and material science. Additionally, the absence of real-time analysis means further exploration is needed to assess performance in time-sensitive scenarios. Implementing these similarity measures in applications such as autonomous navigation or industrial fault detection will require optimization for scalability and responsiveness in dynamic environments.

Future Work: Moving forward, Future research should focus on expanding the dataset to include more varied and complex surface textures, enabling a deeper exploration of how these metrics behave under different conditions and enhancing their generalizability. Integrating machine learning techniques with the similarity metrics could improve their performance in real-time applications, particularly for rapid and accurate correlations between image and sound data. Additionally, studies should address the computational efficiency of these metrics in high-dimensional applications. Exploring their scalability in fields such as robotics, AR/VR, and industrial monitoring will provide valuable insights for practical implementation, ensuring these metrics can be effectively deployed in real-time systems without compromising performance.

This study provides a foundational understanding of how Euclidean Distance, Cosine Similarity, and Pearson Correlation can be effectively used to correlate surface images with audio data generated by movement across those surfaces. Each metric offers unique strengths—Euclidean Distance in capturing texture variations, Cosine Similarity and Pearson Correlation in maintaining stable relationships across diverse textures. Their combined application provides a robust framework for multisensory fusion systems across a range of practical applications.

With further research and optimization, these metrics hold significant promise for improving the accuracy and efficiency of real-time data fusion systems in fields like human-robot interaction, augmented and virtual reality, and industrial monitoring. By addressing the current limitations, such as dataset diversity and real-time performance, future research can advance the field and unlock new possibilities for high-fidelity, multimodal recognition technologies.

References

[1] P. Jiang, C. Kent, and J. Rossiter, "Towards sensory substitution and augmentation: Mapping visual distance to audio and tactile frequency," PLOS ONE, vol. 19, no. 3, p. e0299213, Mar. 2024, doi: 10.1371/journal.pone.0299213.

[2] A. Nakhmani and A. Tannenbaum, "A new distance measure based on generalized Image Normalized Cross-Correlation for robust video tracking and image recognition," Pattern Recognition Letters, vol. 34, no. 3, pp. 315–321, Feb. 2013, doi: 10.1016/j.patrec.2012.10.025.

[3] D. G. Ciric, Z. H. Peric, M. Milenkovic, and N. J. Vucic, "Evaluating Similarity of Spectrogram-like Images of DC Motor Sounds by Pearson Correlation Coefficient," Elektronika ir Elektrotechnika, vol. 28, no. 3, pp. 37–44, Jun. 2022, doi: 10.5755/j02.eie.31041.

[4] V. V. Starovoitov, E. E. Eldarova, and K. T. Iskakov, "Comparative analysis of the SSIM index and the Pearson coefficient as a criterion for image similarity," Eurasian Journal of Mathematical and Computer Applications, vol. 8, no. 1, pp. 76–90, 2020, doi: 10.32523/2306-6172-2020-8-1-76-90.

[5] Y. Zhao, H. Zhao, X. Zhang, and W. Liu, "Vehicle classification based on audio- visual feature fusion with low-quality images and noise," Journal of Intelligent &Fuzzy Systems, vol. 45, no. 5, pp. 8931–8944, Nov. 2023, doi: 10.3233/jifs- 232812.

[6] H. Purwins, B. Li, T. Virtanen, J. Schluter, S.-Y. Chang, and T. Sainath, "Deep Learning for Audio Signal Processing," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 206–219, May 2019, doi: 10.1109/jstsp.2019.2908700.

[7] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fre´chet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms," Inter- speech 2019, Sep. 2019, doi: 10.21437/interspeech.2019-2219.

[8] R. Pethiyagoda, T. J. Moroney, G. J. Macfarlane, and S. W. McCue, "Spectrogram analysis of surface elevation signals due to accelerating ships," Physical Review Fluids, vol. 6, no. 10, Oct. 2021, doi: 10.1103/physrevfluids.6.104803.

[9] Md. I. Ansari and T. Hasan, "SpectNet: End-to-End Audio Sig- nal Classification Using Learnable Spectrograms," arXiv.org, 2022, https://doi.org/10.48550/arXiv.2211.09352.

[10] G. P. Renieblas, A. T. Nogue´s, A. M. Gonza´lez, N. Go´mez-Leon, and E. G. del Castillo, "Structural similarity index family for image quality assessment in radi- ological images," Journal of Medical Imaging, vol. 4, no. 3, p. 035501, Jul. 2017, doi: 10.1117/1.jmi.4.3.035501.

[11] J. Huang, Y. Chen, S. Xiong, and X. Lu, "Cross-Modal Remote Sensing Image–Audio Retrieval With Adaptive Learning for Aligning Correlation," IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–13, 2024, doi: 10.1109/tgrs.2024.3407857.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/tip.2003.819861.

[13] M. Dimbiniaina, D. Pau, and T. A. Naramo, "Mel Power Spectrogram Approximation By Tiny Neural Networks for Home Appliances Classification," in 2023 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT), vol. 32, pp. 60–65, Jun. 2023,doi: 10.1109/metroind4.0iot57462.2023.10180197.

[14] Z. Zheng, J. Chen, X. Zheng, and X. Lu, "Remote Sensing Image Generation From Audio," IEEE Geoscience and Remote Sensing Letters, vol. 18, no. 6, pp. 994–998, Jun. 2021, doi: 10.1109/lgrs.2020.2992324.

[15] J. Zujovic, T. N. Pappas, and D. L. Neuhoff, "Structural Texture Similarity Metrics for Image Analysis and Retrieval," IEEE Transactions on Image Processing, vol. 22, no. 7, pp. 2545–2558, Jul. 2013, doi: 10.1109/tip.2013.2251645.

[16] M. Zelaszczyk and J. Mandziuk, "Audio-to-Image Cross-Modal Generation," in 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, Jul. 2022, doi: 10.1109/ijcnn55064.2022.9892863.

[17] E. Dumic, I. Bacic, and S. Grgic, "Simplified structural similarity measure for image quality evaluation," in Proceedings of the 2012 International Conference on Systems, Signals, and Image Processing (IWSSIP), pp. 442–447, Apr. 2012.

[18] B. Zhang, J. Leitner, and T. Thornton, "Audio recognition using MEL spectro- grams and convolutional neural networks," Noiselab University of California, 2019.

[19] J. Bynum et al., "A Convolutional Neural Network Approach to the Semi-Supervised Acoustic Monitoring of Industrial Facilities," in Proceedings of IC- SIC, Cambridge, UK, pp. 4277–4281, 2019.

[20] S. Qin, H. Shao, Z. Wang, K. Shi, C. Gao, and J. Zhang, "Efficient cosine similarity-based image correlation algorithm for object detection and localiza- tion," in Optoelectronic Imaging and Multimedia Technology IX, vol. 4387, p. 38, Jan. 2023, doi: 10.1117/12.2643921.

[21] X. Qiao, H. Wu, S. K. Roy, and W. Huang, "Hyperspectral Image Classification Based On 3d Sharpened Cosine Similarity Operation," in IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, pp. 7669–7672, Jul. 2023, doi: 10.1109/igarss52108.2023.10281949.

[22] V. Gupta, M. Mittal, V. Mittal, and N. K. Saxena, "Spectrogram as an Emerging Tool in ECG Signal Processing," Recent Advances in Manufacturing, Automation, Design and Energy Technologies, pp. 407–414, Oct. 2021, doi: 10.1007/978-981- 16-4222-7 47.

[23] P. Arun, S. Madhukumar, and P. Careena, "A Method for the Investigation of Bearing Vibration Based on Spectrogram Image Comparison," IOP Conference Series: Materials Science and Engineering, vol. 396, p. 012044, Aug. 2018, doi: 10.1088/1757-899x/396/1/012044.

[24] Z. Wu, B. Zhang, W. Li, and F. Jiang, "A Novel Method for Remaining Use- ful Life Prediction of Bearing Based on Spectrum Image Similarity Measures," Mathematics, vol. 10, no. 13, p. 2209, Jun. 2022, doi: 10.3390/math10132209.

[25] H. Wu¨rschinger and N. Hanenkamp, "Audio Images: Surface Visualization Uti- lizing Impingement Sound of an Air Jet to Detect Tool Wear of Indexable Inserts," SSRN, 2024, doi: 10.2139/ssrn.4697519.

[26] L. Lu¨the, "Calculate similarity: The most relevant metrics in a nutshell," Towards Data Science, accessed 2023.

[27] Thouti, S., Venu, N., Rinku, D. R., Arora, A., & Rajeswaran, N. (2022). Investigation on identify the multiple issues in IoT devices using Convolutional Neural Network. Measurement: sensors, 24, 100509.

[28] Reddy, A. V., Kumar, A. A., Venu, N., & Reddy, R. V. K. (2022). On optimization efficiency of scalability and availability of cloud-based software services using scale rate limiting algorithm. Measurement: Sensors, 24, 100468.

[29] Venu, N., Swathi, R., Sarangi, S. K., Subashini, V., Arulkumar, D., Ralhan, S., & Debtera, B. (2022). Optimization of Hello Message Broadcasting Prediction Model for Stability Analysis. Wireless Communications and Mobile Computing, 2022(1), 2785810.

[30] Venu, N., Revanesh, M., Supriya, M., Talawar, M. B., Asha, A., Isaac, L. D., & Ferede, A. W. (2022). Energy Auditing and Broken Path Identification for Routing in Large-Scale Mobile Networks Using Machine Learning. Wireless Communications and Mobile Computing, 2022(1), 9418172.

[31] Venu, N., Yuvaraj, D., Barnabas Paul Glady, J., Pattnaik, O., Singh, G., Singh, M., & Adigo, A. G. (2022). Execution of Multitarget Node Selection Scheme for Target Position Alteration Monitoring in MANET. Wireless Communications and Mobile Computing, 2022(1), 2088289.

[32] Sujith, A. V. L. N., Swathi, R., Venkatasubramanian, R., Venu, N., Hemalatha, S., George, T., ... & Osman, S. M. (2022). Integrating Nanomaterial and High-Performance Fuzzy-Based Machine Learning Approach for Green Energy Conversion. Journal of Nanomaterials, 2022(1), 5793978