Comparison of Similarity Measures by Implementing a Facial Recognition Biometric System

Wyatt Bender
College of Science and Mathematics
California State University, Fresno
Fresno, USA
wyattbender25@mail.fresnostate.edu

Sarah Hang
College of Science and Mathematics
California State University, Fresno
Fresno, USA
satoji@mail.fresnostate.edu

Amanjot Singh
College of Science and Mathematics
California State University, Fresno
Fresno, USA
aj5in6h@mail.fresnostate.edu

***Abstract*—This report presents a comparative analysis of four similarity measures: Euclidean distance, Cosine distance, Manhattan distance, and Chebyshev distance through the implementation of a facial recognition biometric system. By determining the accuracy and computational efficiency of these measures, using a dataset consisting of 40 celebrity faces, we aim to identify the most effective distance measure for face recognition. Through extensive experimentation and evaluation, we discover that Cosine distance appears to be the most effective measure, due to the high performance in terms of finding true positives and minimizing false positives. Our findings suggest that the choice of distance measure significantly impacts the accuracy and reliability of facial recognition systems, with potential implications for broader biometric applications. Additionally, we discuss opportunities for future research, including the exploration of multimodal biometric systems and the adaptation of distance measures to other biometrics.**

***Index Terms*—Similarity measures, Facial Recognition, Biometric Systems, Euclidean distance, Cosine similarity, Manhattan distance, Chebyshev distance**

I. Introduction

With the various similarity measures used in biometric systems, we perform a comparison of similarity measures: Euclidean distance, Cosine similarity, Manhattan distance, and Chebyshev distance by implementing a facial recognition biometric system. We measure the true positives, true negatives, false positives, and false negatives of each similarity measure by comparing the accuracy and time it takes to compare 40 celebrities from a dataset. By identifying the accuracy and length of time it takes for each similarity measure to do it's job, we have a better understanding of each distance measure and how it works with face embeddings. and how they perform in a facial recognition biometric system. The goal of this paper is to identify which similarity measures outperform each other by finding the most true positives and true negatives while reducing the false positives and false negatives.

- True positive: the input user is a registered user and matches with a template user.
- True negative: the input user is not a registered user and does not match with any template users.
- False positive: the input user is not a registered user and matches with template user.
- False negative: the input user is a registered user but does not match with any template users.

II. Related Work and Background Material

We first analyzed two research papers before attempting our project. The first paper, "Feature Selection Using Euclidean Distance and Cosine Similarity for Intrusion Detection Model" [1] discusses intrusion detection system (IDS) and how feature selection is a critical step of data mining. The paper categorizes intrusion detection techniques into anomaly detection and misuse detection. Despite the success of intrusion detection systems in detecting known attacks, they still struggle with detecting unknown attacks, leading to false negatives. To address this issue, the paper proposes a feature selection method using Euclidean Distance and Cosine Similarity. Euclidean Distance is utilized for selecting features to detect known attacks, while Cosine Similarity is used for detecting unknown attacks. It suggests that feature selection based on Euclidean Distance and Cosine Similarity can enhance the accuracy and efficiency of IDSs in detecting known and unknown attacks. While the focus of our project is to compare similarity measures for facial recognition, we have a better understanding of how Euclidean distance and Cosine distance can be used for various kinds of feature vectors.

The second paper, "Distance based verification techniques for online signature verification system" [2] discusses verification of online signatures using Manhattan distance, Chebyshev distance, and Euclidean distance. This paper utilizes histogram feature extraction to extract the features of signatures for the signatures to be represented as a feature vector. The findings of the paper reveals that Euclidean distance resulted in the most accurate compared to Manhattan and Chebyshev based on the calculated equal error rate (EER). The EER is the rate at which the false acceptance rate (FAR) and false rejection rate (FRR) are equal. Minimizing the EER provides a better threshold value to obtain the most ideal threshold at each distance measure. This paper gives us a better understanding of the

effectiveness of Euclidean distance, Manhattan distance, and Chebyshev distance with the feature extraction of a signature since our project utilizes all three of these distance measures along with Cosine distance for face embeddings.

III. Methods and Implementation

A proper dataset and the features extracted from the dataset plays a crucial role in the process of comparing distance measures when implementing a facial recognition biometric system.

A. Dataset

We found two datasets on Kaggle called Face Recognition [3] and 14 Celebrity Faces [4]. We combined the two datasets together and also added one to two additional faces found online. We needed a dataset that involved several images of one person so that we can use the same face for comparison, and since there are many images of celebrities throughout on the web, we decided to utilize a dataset of celebrity faces. Both datasets contain anywhere between 20 to 50 images of a celebrity and we use a total of 40 celebrities faces for the purpose of the facial recognition biometric system. Then, we choose 30 celebrities and register their faces as templates. A different image of the 30 registered celebrities were then used along with 10 unregistered celebrities, so that we can compare a total of 40 faces to all 30 templates.

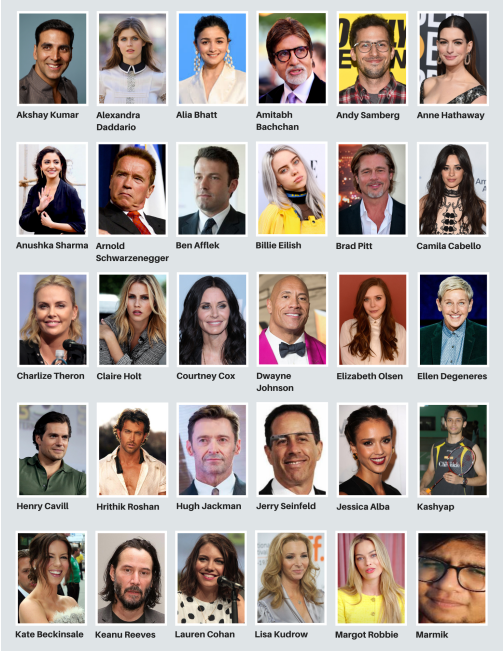

Fig. 1: Dataset of registered faces

B. Feature Extraction

Feature extraction plays a crucial step in computing distance measures. Feature extraction is performed by preprocessing a face. The model takes an image as input, applies several convolutional and max-pooling layers to extract features, flattens the output, passes it through fully connected layers for more processing and outputs a feature vector of length 128 at the last layer. The preprocessing function loads and image, resizes the image to 224x224 pixels, converts it into an array and expands its dimensions to make it work with the model's input shape. The feature extraction function takes in an array of image paths and the feature extraction model as parameters. It preprocesses the face and passes the preprocessed image through the feature extraction model to obtain a feature vector representing the image. The feature vector is then essentially returned and is used to compute the distance measures between two faces.

IV. Distance Measures

The four similarity measures we compare include Euclidean distance, Cosine distance, Manhattan distance, and Chebyshev distance. These four distance measures were four out of many that were reliable to work with our feature vectors. It is important to note that a threshold plays an important role in calculating distances to properly be classified as a true positive, false positive, false negative, and true negative. For a fair comparison between all distance measures, we normalized the thresholds to 0%, 25%, 50%, 75%, and 100%, computed the distances for all 1,200 face comparisons and obtained the minimum and maximum value for each distance measure. We then computed a normalized threshold for each distance measure by subtracting the distance measure's respective maximum and minimum values together, multiplying the percentage, and adding the minimum value. The normalized thresholds then gave us a accurate comparison to determine which distance measures worked better given our dataset and feature vectors.

A. Euclidean Distance

The Euclidean distance is a measure of the straight line distance between two points similar to the Pythagorean theorem. The formula to calculate Euclidean distance between two feature vectors a and b is

$$distance = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{1}$$

- a_i and b_i are the components of the feature vectors.
- n is the dimensionality of the feature vectors.

The normalized threshold for Euclidean distance at 50% was 60.714 with the minimum values being 30.103 and a maximum value of 91.325 when computing all the feature vectors using Euclidean distance. Thus, the Euclidean distance is used as a measure of dissimilarity between feature vectors representing faces. If the Euclidean distance between the feature vectors of two faces falls below a certain threshold, they are considered to be the same person. Otherwise, they are considered different individuals.

B. Cosine Distance

Cosine distance is a measure of dissimilarity between two vectors by calculating the angular difference between the two vectors. Cosine distance is particularly useful when considering the direction of vectors instead of their absolute values. The cosine distance is calculated based on the dot product

of the normalized feature vectors. The dot product of two vectors a and b is computed as the sum of the products of their corresponding components:

$$\text{dot product} = \sum_{i=1}^{n} a_i \cdot b_i \tag{2}$$

- a_i and b_i are the components of the feature vectors.

Cosine similarity, which measures the cosine of the angle between the two vectors, is calculated using the dot product and the magnitudes of the vectors:

$$\text{cosine similarity} = \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\|\mathbf{a}\|\|\mathbf{b}\|} \tag{3}$$

- $\|\mathbf{a}\|$ is the magnitude of vector $\mathbf{a}$.
- $\|\mathbf{b}\|$ is the magnitude of vector $\mathbf{b}$.

Cosine distance is computed by subtracting the cosine similarity from 1. This step ensures that larger values of cosine similarity (indicating greater similarity between vectors) result in smaller cosine distances.

$$\text{cosine distance} = 1 - \text{cosine similarity} \tag{4}$$

Cosine distance measures the dissimilarity between two vectors by comparing the angle between them. It ranges from 0 (meaning perfect similarity) to 2 (meaning complete dissimilarity), with larger values indicating greater dissimilarity.

C. Manhattan Distance

Manhattan distance is a measure of the distance between two points in a grid. Manhattan distance is calculated by subtracting the corresponding values of the two feature vectors, obtaining the absolute value of the difference and then adding up all of the values together to get the total Manhattan distance. The formula to calculate Manhattan distance between two feature vectors a and b is

$$\text{Manhattan distance} = \sum_{i=1}^{n} |a_i - b_i| \tag{5}$$

- a_i and b_i are the components of the feature vectors.
- n is the dimensionality of the feature vectors.

Unlike Euclidean distance, which measures the straight-line distance between two points, Manhattan distance accounts for movements along axes and is much simpler to implement.

D. Chebyshev Distance

Chebyshev distance is a metric used to measure the maximum difference between corresponding feature vectors. Chebyshev distance is useful when considering the maximum values between two vectors, regardless of their individual positions. The formula for Chebyshev distance is quite similar to Manhattan distance except that it takes the maximum computed value as the distance. The formula to calculate Chebyshev distance between two feature vectors is

$$d = \max(|x_2 - x_1|, |y_2 - y_1|) \tag{6}$$

- a_i and b_i are the components of the feature vectors.

The absolute differences between corresponding components are calculated, and then the maximum of these absolute differences is taken as the Chebyshev distance. The resulting Chebyshev distance represents the maximum difference between any corresponding elements of the two vectors.

E. Other Distance Measures

Various distance measures can be used in biometric systems as shown in table 1. The choice of distance measures depends on the types and characteristics of the data used. Several distance measures have different calculations and measurements between two feature vectors including the angle, straight-line distance, binary data, correlations, and sets.

TABLE I: Other Various Distance Measures

Distance Measures	***Description***
Minkowski Distance	Implementation of Euclidean and Manhattan distances, where p determines the type of distance.
Mahalanobis Distance	Measures the distance between a point and a distribution.
Hamming Distance	Measures the count of how many corresponding symbols of two vectors are different.
Jaccard Distance	Measures similarity between sets by computing the intersection divided by the cardinality of the union of the two sets.
Levenshtein Distance	Measures the minimum edit distance between strings.
Canberra Distance	Measures the distance between two pairs of points.
Haversine Distance	Calculates distance between latitude and longitude.
Correlation Distance	Measures dissimilarity based on correlation.
Bhattacharyya Distance	Measures similarity between probability distributions.
Dice Distance	Measures similarity between sets.

V. Results

The results are in a confusion matrix shown in figure 2 containing the true positives, true negatives, false positives, and false negatives for each distance measure with a normalized threshold of 50%. The figure shows that Euclidean distance and Manhattan distance results are the same despite their different calculations and outputs at the same threshold. Chebyshev performed really well at a 50% threshold but also had a somewhat small number of false positives with a larger number of true negatives. Cosine distance had the lowest number of true positives but it did also have the lowest number of false positives which shows that Cosine distance may be more ideal for this particular dataset with the proper adjusted threshold. Typically when selecting a threshold for a distance measure in any biometric system, it is important to select one that balances the system's accuracy in verifying or identifying individuals.

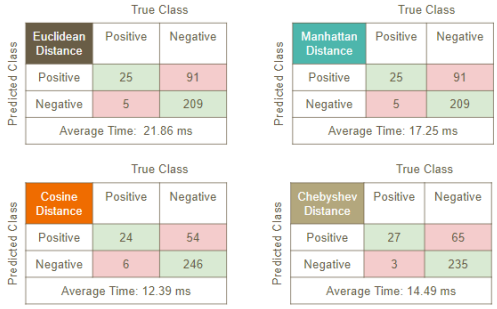

Fig. 2: Results in a Confusion Matrix

The false acceptance rate (FAR) and false rejection rate (FRR) of each distance measure is graphed as shown in figure 3. The thresholds used for each measure were normalized to 0%, 25%, 50%, 75%, and 100% in order to graph the FAR and FRR. The FAR is the rate where the system incorrectly accepts imposters as genuine matches and the FRR is the rate where the system incorrectly rejects genuine matches. The graph is similar to the results seen in Figure 2 where the distance measures have very similar trends at a 50% threshold. Though, since the thresholds were normalized at percentages the graphs shown in figure 3 essentially look very similar with similar trends.

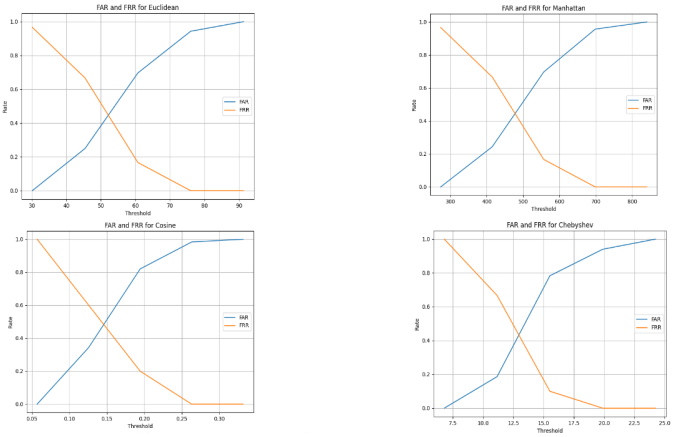

Fig. 3: FAR and FRR of All Distance Measures

Euclidean distance and Manhattan distance appear to be simple and reliable but were shown to be the worst overall when comparing them at a 50% threshold. Chebyshev distance identified a lot of true positives early on at 50% outperforming Euclidean distance and Manhattan distance. Cosine distance appeared to be the best even at a 50% threshold due to having the least amount of false positives and a good amount of true positives. Overall Cosine distance appeared to perform better for our data in our facial recognition biometric system.

VI. Conclusion

In this project, we conducted an analysis of various similarity measures of facial recognition biometric systems. By evaluating Euclidean distance, Cosine distance, Manhattan distance, and Chebyshev distance, we wanted to find the most effective measure for accurately identifying individuals while minimizing false identifications. By utilizing a dataset consisting of 40 celebrity faces, we measured the true positives, true negatives, false positives, and false negatives, discovering the performance and efficiency of each measure.

We wanted to find importance of similarity measures used in biometric systems and identify the most suitable measure for facial recognition biometric systems. Drawing from related research in intrusion detection and online signature verification systems, we utilized the results of feature selection and distance-based verification techniques to understand the various distance measures used throughout other biometric systems.

Feature extraction played a major role in the project by utilizing our model and obtaining a feature vector at the last layer to compute distance measures for 1,200 face embeddings.

For future work, we could consider applying the distance measures to different biometrics such as fingerprints or palm prints. It would also be possible to incorporate more users, both registered and unregistered users, for more comparisons. Additionally, it would be effective to combine distance measures for a multimodal biometric system by combining distance measures together such as Euclidean and Cosine so that a face would have to pass through two distance measures to be classified as a matching face.

In conclusion, the distance measures that are being used in general depends on the feature vectors, threshold, and types of data used. Our project provides a better understanding of distance measures in biometric systems and emphasizes their large role in ensuring the accuracy and reliability of facial recognition technologies. By discovering the strengths and weaknesses of various distance measures, we provide a foundation for the consistent use of facial recognition biometric systems.

References

[1] A. Suebsing and N. Hiransakolwong, "Feature Selection Using Euclidean Distance and Cosine Similarity for Intrusion Detection Model," 2009 First Asian Conference on Intelligent Information and Database Systems, Dong hoi, Vietnam, 2009, pp. 86-91, doi: 10.1109/ACIIDS.2009.23. keywords: Euclidean distance;Intrusion detection;Data mining;Computer networks;Machine learning;Mathematics;Computer science;Robustness;Feature extraction;Deductive databases;Intrusion Detection System (IDS);Feature Selection;Data Mining;Cosine Similarity and Euclidean Distance,

[2] M. Arora, H. Singh and A. Kaur, "Distance based verification techniques for online signature verification system," 2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS), Chandigarh, India, 2015, pp. 1-5, doi: 10.1109/RAECS.2015.7453285. keywords: Handwriting recognition;Histograms;Chebyshev approximation;Euclidean distance;Feature extraction;Error analysis;Hidden Markov models;Online signature verification;feature extraction;histogram;Manhattan;Chebyshev;Euclidean distance;EER,

[3] S. Singh, "Face recognition dataset," Kaggle, https://www.kaggle.com/datasets/cybersimar08/face-recognition-dataset (accessed May 3, 2024).

[4] Danupnelson, "14 celebrity faces dataset," Kaggle, https://www.kaggle.com/datasets/danupnelson/14-celebrity-faces-dataset (accessed May 3, 2024).