A Scalable and Efficient Similarity Search Framework for Large-Scale Face Recognition Systems

Peethani Satya Durga Rao[1], Y. Dayanand Kumar[2]
[1]*M. Sc Artificial Intelligence and Data Science, Central University of Andhra Pradesh, Anantapur, India*
[2]*Assistant Professor, Department of Computer Science & AI, Central University of Andhra Pradesh, Anantapur, India*

***Abstract*—These instructions give you guidelines for preparing papers for the International Journal of Innovative Research in Technology (IJIRT). Recent advances in deep learning have significantly improved the accuracy of face recognition systems through the use of facial embeddings; however, scalability remains a critical challenge in large-scale deployments. As the number of registered identities increases, searching within high-dimensional feature spaces becomes computationally expensive, reducing system efficiency. This study investigates the performance and scalability of multiple similarity search techniques for large-scale face recognition. Facial embeddings are generated using the FaceNet deep neural network, producing 128-dimensional feature vectors. Five search methods Linear Search, KD-Tree, Ball-Tree, FAISS Exact Search, and FAISS with Inverted File (IVF) indexing are evaluated using the Labeled Faces in the Wild (LFW) dataset, with database sizes expanded from 5,000 to 40,000 images. Experimental results show that traditional methods, including Linear Search and tree-based structures, experience significant performance degradation as dataset size increases due to computational complexity and high-dimensional data challenges. In contrast, FAISS-based methods, particularly IVF indexing, achieve substantially faster search times while maintaining strong recognition performance. The system achieves an Area Under the Curve (AUC) of approximately 0.974 and an Equal Error Rate (EER) of 0.07, demonstrating reliable discriminative capability. A limitation of this study is the use of dataset replication to simulate large-scale conditions, which may not fully capture real-world data diversity. Overall, the findings highlight the effectiveness of vector indexing techniques for building scalable and efficient face recognition systems.**

***Index Terms*—Face Recognition, Similarity Search, FAISS, Facial Embeddings, Biometric Systems**

I. INTRODUCTION

There has been a marked improvement in the performance of the face recognition systems with the implementation of the facial embedding technology through deep learning methods; however, one major drawback in the system has been its lack of scalability. This limitation is associated with an increased complexity in searching for the face features that reside in a higher dimensional space as the number of enrolled individuals increases. Hence, this paper will evaluate the efficiency and scalability of the similarity search techniques in large scale face recognition systems.

The generation and storage of the facial embeddings will be done using the deep neural network known as FaceNet. The facial embeddings will have dimensions of 128 while the five search approaches to be considered include Linear Search, KD-tree, Ball-tree, FAISS exact search and FAISS Inverted file (IVF) indexing. The test will be performed on the Labeled Faces in the Wild (LFW) datasets by increasing the dimension of the database from 5,000 up to 40,000.

The progress made in deep learning has contributed significantly to improving the efficiency of face recognition algorithms. Specifically, convolutional neural networks can learn complex features from facial images automatically. Methods like FaceNet represent faces as vectors in a low-dimensional embedding space, such that images of the same individual will be closely located, whereas images of different people will be far apart from each other [1]. Moreover, there are approaches such as ArcFace that improve the ability of representing the input faces using an angular distance metric, resulting in better accuracy on benchmarks [2].

Despite the progress made in terms of feature representation technologies, there are still many difficulties when implementing face recognition systems at a scale that would be applicable in the real world. As the number of registered users grows from several thousand individuals up to several million people, searches in high-dimensional space become very computationally expensive. In this case, finding the most similar image among all available options becomes a bottleneck because of the huge search space. Traditional solutions such as Linear Search involve the comparison of the query with each individual image which entails linear computational complexity that cannot be used in practical implementations.

In order to increase the speed of the search, spatial indexes like KD-Tree and Ball-Tree are used. They are indeed faster but their efficiency decreases as the number of dimensions increases due to the curse of dimensionality, making them not efficient enough for large-scale applications. More recently, vector search libraries such as Facebook AI Similarity Search (FAISS) have been introduced to support efficient retrieval in high-dimensional feature spaces [3]. These approaches rely on advanced indexing mechanisms, including clustering techniques and approximate nearest neighbor strategies, to lower computational overhead while preserving retrieval accuracy.

Problem Statement

- Scalability in large-scale face recognition systems is hindered by the computational cost of high-dimensional similarity searches.
- Traditional nearest neighbor search methods suffer from linear computational complexity, making them unviable for large databases.
- The curse of dimensionality severely limits the effectiveness of standard spatial indexing methods like KD-Trees and Ball-Trees for 128-dimensional facial embeddings.

Objectives

- To evaluate the scalability and performance of various similarity search techniques, including Linear Search, KD-Tree, Ball-Tree, and FAISS indexing structures.
- To analyze computational efficiency by progressively scaling the facial embedding database up to 40,000 vectors.
- To assess the trade-offs between search speed and biometric recognition reliability using standard metrics such as ROC, AUC, and EER.

Building on these developments, this study examines the scalability and performance of different similarity search methods within the context of large-scale face recognition. Five techniques Linear Search, KD-Tree, Ball-Tree, FAISS Exact Search, and FAISS with Inverted File (IVF) indexing are comparatively analyzed across datasets of increasing size constructed from the LFW dataset [4].

II. RELATED WORK

Face recognition has been an active research area in computer vision for several decades. Early face recognition systems primarily relied on handcrafted feature extraction techniques combined with statistical pattern recognition methods. One of the earliest approaches is the Eigenfaces method proposed by Turk and Pentland, which used Principal Component Analysis (PCA) to represent facial images in a lower-dimensional subspace [5]. To improve the discriminative power of face representations, Fisherfaces were later introduced using Linear Discriminant Analysis (LDA) [6].

The introduction of deep learning significantly improved the robustness and accuracy of face recognition systems. Convolutional Neural Networks (CNNs) are capable of automatically learning hierarchical feature representations directly from raw image data [7]. Another major breakthrough in face recognition research was the FaceNet model introduced by Schroff et al. [1]. FaceNet learns to map facial images directly into a compact Euclidean embedding space where images belonging to the same individual are located close together while images of different individuals are separated by larger distances. Subsequent research further improved embedding quality using advanced loss functions [2].

While deep learning models provide highly discriminative facial embeddings, the efficiency of searching these embeddings in large databases remains a critical challenge. Traditional nearest neighbor search methods such as Linear Search perform exhaustive comparisons and therefore become computationally expensive when the number of stored embeddings grows. To address this limitation, tree-based indexing methods such as KD-Tree and Ball-Tree have been used to accelerate nearest neighbor search in multi-dimensional spaces [8]. However, the performance of these tree-based structures often deteriorates when dealing with very high-dimensional feature vectors due to the curse of dimensionality.

As a result, researchers have explored approximate nearest neighbor (ANN) search techniques to improve scalability [9]. More recently, Facebook AI Research introduced FAISS (Facebook AI Similarity Search), a high-performance library designed for efficient similarity search and clustering of dense vectors [3]. FAISS supports both exact search and approximate nearest neighbor search using optimized indexing structures such as inverted file systems and product quantization.

III. PROPOSED FRAMEWORK

This study proposes a scalable similarity search framework for large-scale face recognition systems. The framework integrates deep facial embedding generation with efficient nearest neighbor search algorithms in order to improve the computational efficiency of identity matching in large biometric databases. Modern face recognition systems typically represent facial images as numerical feature vectors called embeddings [1].

The proposed framework consists of four main stages: image preprocessing, embedding extraction, similarity search indexing, and identity prediction. Initially, facial images are obtained from the dataset and processed through face detection and alignment operations to normalize the input data. After preprocessing, deep neural networks are used to extract facial embeddings. In this research, the FaceNet model is used to generate 128-dimensional embedding vectors.

Once embeddings are generated, they are stored in a searchable embedding database. The similarity search stage then identifies the closest matching embedding to the query vector. Several search algorithms can be applied for this task, including brute-force search, tree-based indexing structures, and approximate nearest neighbor methods. To overcome the limitations of traditional tree structures, the proposed framework incorporates FAISS-based indexing techniques [3].

[Insert Figure 1: Architecture of the proposed scalable face recognition framework]

IV. EXPERIMENTAL SETUP

This section describes the experimental configuration used to evaluate the proposed similarity search framework for large-scale face recognition.

A. Dataset

The experiments were conducted using the Labeled Faces in the Wild (LFW) dataset, which is widely used as a benchmark dataset for face recognition research. Identities with only a single image were removed during preprocessing.

B. Embedding Generation

Facial embeddings were generated using the FaceNet deep learning model [1]. Each face image was transformed into a 128-dimensional embedding vector.

C. Similarity Search Algorithms

To evaluate the scalability of different search strategies, five similarity search algorithms were implemented: Linear Search, KD-Tree, Ball-Tree, FAISS Exact Search, and FAISS IVF Search.

D. Dataset Scaling

Because the filtered LFW dataset contains a limited number of images, the embedding database was synthetically expanded using replication techniques to simulate datasets of increasing scale: 5,000, 10,000, 20,000, and 40,000 embeddings.

E. Evaluation Metrics

Recognition performance was evaluated using accuracy, precision, recall, and F1-score. Receiver Operating Characteristic (ROC) analysis was performed to analyze the trade-off between true positive rate and false positive rate. For biometric system evaluation, the Equal Error Rate (EER) was also computed.

V. RESULTS AND DISCUSSION

This section presents the experimental results and a comparative evaluation of the five similarity search algorithms.

A. Scalability and Search Speed Analysis

Table I Search Time Comparison Across Different Dataset Sizes (in seconds)

Dataset Size	Linear Search	KD-Tree	Ball-Tree	FAISS Exact	FAISS IVF
5000	0.1250	0.1352	0.1126	0.0032	0.0012
10000	0.2445	0.2725	0.2445	0.0060	0.0022
20000	0.5940	0.5557	0.5270	0.0133	0.0040
40000	1.0825	0.9843	1.3044	0.0257	0.0065

As observed in Table I and Figure 2, Linear Search exhibits a steep, linear increase in computation time as the dataset grows. Tree-based methods (KD-Tree and Ball-Tree) provide only marginal improvements over Linear Search. Their efficiency heavily deteriorates when handling 128-dimensional embedding vectors due to the curse of dimensionality.

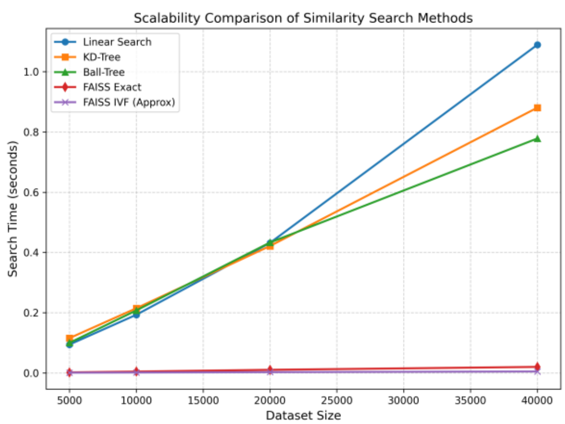

Fig. 2. Scalability comparison showing dataset size versus search time.

In stark contrast, FAISS-based approaches maintain significantly lower search times. Figure 3 highlights the relative speedup achieved by vector indexing. FAISS IVF proves to be the most efficient.

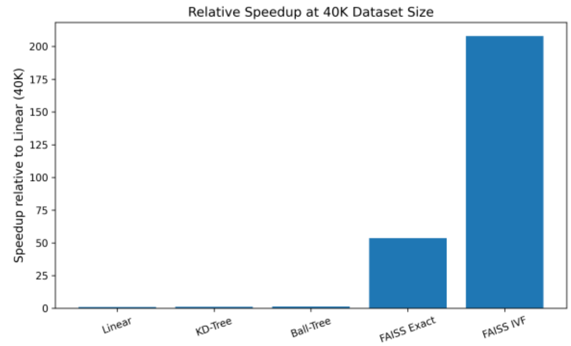

Fig. 3. Speed comparison between similarity search algorithms.

B. Recognition Performance

Because all search methods evaluated in this study retrieve matches from the identical set of FaceNet embeddings, the underlying recognition accuracy of the system remains consistent across all algorithms.

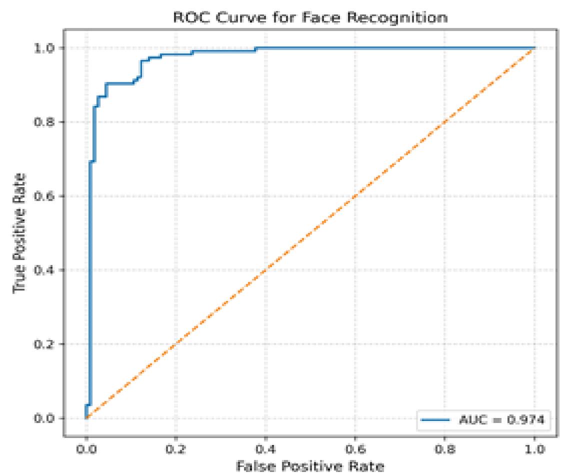

Fig. 4. Receiver Operating Characteristic (ROC) curve for the face recognition system.

The ROC curve (Figure 4) illustrates the relationship between the true positive rate and the false positive rate across different threshold values. The system achieved an Area Under the Curve (AUC) of approximately 0.974.

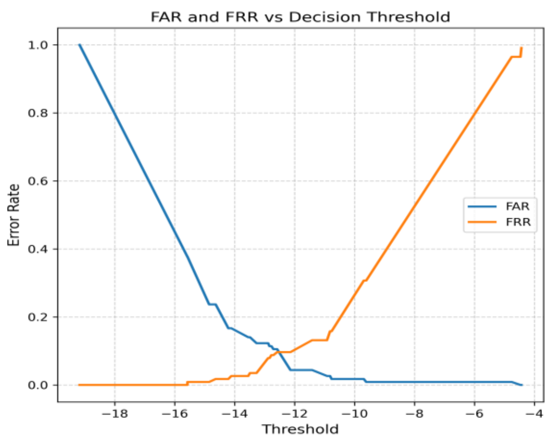

Fig. 5. FAR-FRR curve used for Equal Error Rate analysis.

Furthermore, Figure 5 presents the False Acceptance Rate (FAR) and False Rejection Rate (FRR) curves. The intersection of these curves yields an Equal Error Rate (EER) of approximately 0.07.

VI. CONCLUSION

This research investigated the scalability of similarity search algorithms for face recognition systems based on deep facial embeddings. The experimental results showed that traditional search methods such as Linear Search experience significant performance degradation as the number of stored embeddings increases. In contrast, FAISS-based indexing techniques demonstrated substantially better scalability and computational efficiency.

The FAISS IVF method achieved the fastest search time among the evaluated algorithms. The experiments achieved an AUC value of approximately 0.974 and an Equal Error Rate of approximately 0.07, indicating strong classification capability and reliable identity discrimination [10]. Overall, the findings confirm that scalable similarity search frameworks play a crucial role in enabling efficient face recognition systems.

ACKNOWLEDGMENT

P. S. D. Rao thanks Y. Dayanand Kumar and the Ashok Cuap Faculty for their continuous guidance, meticulous feedback, and unwavering support throughout the development of this dissertation and research framework.

REFERENCES

[1] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.

[2] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4690–4699.

[3] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” IEEE Transactions on Big Data, 2017.

[4] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” Univ. Massachusetts, Amherst, MA, USA, Tech. Rep., 2007.

[5] M. Turk and A. Pentland, “Eigenfaces for recognition,” Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71–86, 1991.

[6] P. Belhumeur, J. Hespanha, and D. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711–720, 1997.

[7] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1701–1708.

[8] J. L. Bentley, “Multidimensional binary search trees used for associative searching,” Communications of the ACM, vol. 18, no. 9, pp. 509–517, 1975.

[9] M. Muja and D. G. Lowe, “Fast approximate nearest neighbors with automatic algorithm configuration,” in Proc. Int. Conf. Computer Vision Theory and Applications (VISAPP), 2009, pp. 331–340.

[10] A. K. Jain, A. A. Ross, and K. Nandakumar, Introduction to Biometrics. Springer, 2011.