IBB Research Track: Multimodal person recognition

Sebastijan Trojer
IBB 2024/25 , FRI, UL
st5804@student.uni-lj.si

***Abstract*—This work investigates multimodal biometric recognition by integrating face, iris, and fingerprint data to enhance accuracy and robustness. We compare the performance of late fusion with traditional unimodal approaches, highlighting how multimodal fusion consistently outperforms unimodal methods. A combined dataset was created by merging the CelebFaces, Casia-Iris-Thousand, and Casia Fingerprint datasets, addressing demographic and sample imbalances via subsampling. Our experiments demonstrate that k-Nearest Neighbors (kNN), combined with PCA for dimensionality reduction, excels in modality fusion, achieving perfect classification accuracy with the full modality set. These results underscore the effectiveness of simple, interpretable multimodal fusion techniques in biometric recognition tasks.**

I. Introduction

Multimodal learning leverages diverse data sources to extract complementary information, often outperforming single-modality approaches. Combining modalities enhances representations and learning outcomes, particularly in classification and prediction tasks. This work highlights the benefits of multimodal data training, demonstrating the advantages over unimodal methods.

We focus on neural networks for feature extraction, complemented by simpler, effective models for final decision-making. Although advanced fusion techniques such as EmbraceNet were considered and briefly evaluated, the complexity of such specialized models often leads to overfitting and marginal improvements that do not justify their lack of simplicity and explainability. In contrast, predictions based on late fusion from individual modalities proved both effective and interpretable, outperforming complex models in practical scenarios.

In this work, we address the task of person recognition by integrating three biometric modalities: iris, face, and fingerprint images. By combining these modalities, we aimed to improve recognition accuracy and robustness, particularly in scenarios where one or more modalities may be unreliable. This multimodal approach highlights the advantages of fusing diverse biometric data for more reliable and secure person recognition systems.

II. Related Work

Lahat et al. [1] define a modality as data captured by a sensor and multimodal learning as the process of training on data from multiple sensors that observe the same phenomenon. The primary motivation for multimodal learning lies in its ability to extract complementary information from different modalities, providing richer insights and improving model performance. Techniques for data fusion, a key component of multimodal learning, have been extensively researched. Early (data-level) fusion and late (decision-level) fusion are two general categories, with the latter gaining popularity through the rise of ensemble classifiers [2]. Ramachandram et al. [3] offer an in-depth analysis of these techniques, including those based on neural networks.

Bayoudh et al. [4], in their survey on deep multimodal learning for computer vision, explore a wide range of modality fusion techniques. These approaches include classical methods from the early 2000s, such as hidden Markov model-based fusion [5], canonical correlation analysis [6], and deep belief networks [7]. They also cover more recent advancements, including step-based deep multimodal autoencoders [8] and bimodal convolutional neural networks [9]. Furthermore, Peng et al. [10] conducted a comprehensive survey on transformer-based multimodal learning, emphasizing the growing interest and sophistication in this area.

Among specialized fusion techniques, EmbraceNet [11] stands out as a robust approach for handling multimodal data. The method is particularly noted for its ability to manage scenarios where data from some modalities is partially or entirely missing. In their study, Choi et al. compared EmbraceNet to several fusion methods, including early and late fusion, multimodal autoencoders, and compact multilinear pooling. EmbraceNet consistently outperformed these methods in terms of robustness and accuracy, even with incomplete data.

III. Methodology

A. Modality fusion

Modality fusion refers to integrating diverse data types to improve the performance of machine learning models. It combines the strengths of multiple data sources, improving upon single-modality approaches. Generally, it is divided into three categories: early, late, and hybrid fusion [4]:

- Early Fusion: Combines raw data features from multiple modalities before model training. This enables the model to learn joint representations, capturing interactions between modalities. It is effective in tasks like audiovisual speech recognition, where synchronized feature extraction improves performance in noisy environments.
- Late Fusion: Processes each modality with specialized models, combining their outputs during the decision phase. This approach is flexible and allows independent fine-tuning of models, making it suitable for modalities with differing characteristics.
- Hybrid Fusion: Merges elements of early and late fusion, combining multimodal features at intermediate stages for greater adaptability.

For this task we decided to also compare simpler techniques, namely late fusion, to more advanced ones, specifically EmbraceNet [11]. EmbraceNet is a model consisting of two types of layers:

- Docking layers: Converts feature vectors from all modalities into uniform shapes for integration.
- Embracement layer: Joins the vectors from the docking layers into a single vector based on multinomial sampling method.

The resulting feature vector is then passed to an exit classifier for the final prediction. This architecture aims to handle partial modality availability effectively while providing robust multimodal learning capabilities.

B. Feature extraction

In our experiments we used the following models: EfficientNet [12] and FaceNet [13].

EfficientNet is an architecture proposed by Tan and Le [12], where they study how to effectively scale convolutional neural networks while maintaining model efficiency. The proposed architecture surpasses the state-of-the-art accuracy on ImageNet and five other commonly used transfer learning datasets, with significantly less parameters and floating point operations per second (FLOPS).

FaceNet is an older, specialized model, proposed by Schroff et al. [13], where a new type of loss function is used to train the model. Instead of optimizing the final network output, the introduced triplet loss focuses on optimizing the embedding itself, resulting in greater representational efficiency.

IV. Experiments

A. Data

Due to the absence of a multimodal biometric dataset that suited our requirements, we created a combined dataset by integrating three distinct datasets. Each dataset provided a unique modality, as detailed in Table I. An example three-tuple of a subject sample from the combined dataset is shown if Figure 1.

Dataset	Modality	# Subjects	# Samples
CelebFaces [14]	Face	10 177	202 599
Casia-Iris-Thousand [15]	Iris	1000	20 000
Casia Fingerprint v5 [15]	Fingerprint	500	20 000
Combined Dataset	Multimodal	500	42 011

TABLE I
Dataset Characteristics

Fig. 1. Example of the samples belonging to a subject in the combined dataset.

The data was combined based on subjects: For each subject in CelebFaces, two subjects were randomly selected from the Casia datasets, and their data was aggregated under a new subject ID in the combined dataset. While the Casia datasets had a consistent number of samples per subject, CelebFaces occasionally had fewer samples. This imbalance was mitigated through oversampling or subsampling to ensure uniformity and only results using subsampling are reported. Additionally, it is important to note that the Casia datasets were collected primarily from East Asian demographics, while CelebFaces consists mainly of Western demographic data. This demographic disparity may introduce variations in feature distributions, which could impact the performance and generalization of the models.

B. Classification pipeline

The classification pipeline consisted of the following key steps:

1) **Feature Extraction**: EfficientNet and FaceNet were used for modality feature extraction. EfficientNet was applied to the Casia datasets (iris and fingerprint), achieving 95% and 86% validation accuracy for iris and fingerprint data, respectively. FaceNet was used for the CelebFaces dataset (face modality), leveraging a pretrained model for feature extraction. The models were trained on 50% of each subject's data, with the remaining data used for evaluation.
2) **Modality Fusion**: Late fusion was employed to generate two versions of each modality combination:

Modality Set	Preprocessing	Classifier	Accuracy	F1 Score
All	/	Dummy Classifier	0.002	0.00
Face	/	Random Forest	0.23	0.21
Iris	/	Random Forest	0.74	0.70
Fingerprint	/	Random Forest	0.69	0.67
Face	PCA	Random Forest	0.36	0.31
Iris	PCA	Random Forest	0.89	0.87
Fingerprint	PCA	Random Forest	0.84	0.83
Face + Iris	PCA	Random Forest	0.95	0.94
Face + Fingerprint	PCA	Random Forest	0.93	0.91
Iris + Fingerprint	PCA	Random Forest	0.95	0.94
All	PCA	Random Forest	0.98	0.97
Face	**PCA**	**kNN**	**0.53**	**0.48**
Iris	**PCA**	**kNN**	**0.93**	**0.92**
Fingerprint	**PCA**	**kNN**	**0.93**	**0.92**
Face + Iris	**PCA**	**kNN**	**0.99**	**0.99**
Face + Fingerprint	**PCA**	**kNN**	**0.97**	**0.97**
Iris + Fingerprint	**PCA**	**kNN**	**0.99**	**0.99**
All	**PCA**	**kNN**	**1.00**	**1.00**

TABLE II
Classification Results for Various Modality Sets and Classifiers. The dummy classifier row serves as a baseline. We report results for individual modalities, modality combinations, and preprocessing techniques like PCA. The combinations with best performance are highlighted in bold.

one where the more abundant modalities were subsampled, and another where the underrepresented modalities were oversampled. To avoid potential overfitting associated with oversampling, only the results from the undersampled data are reported.

3) **Recognition**: For recognition, we utilized lightweight machine learning models to evaluate the performance across different modality combinations. 80% of the data remaining from feature extraction was used for training and 20% for evaluation.

Since the embeddings from EfficentNet were consisting of over 1400 features, we also tested prediction performance on the reduced versions of the embeddings, which we acquired using PCA to reduce the dimensionality to 100 components.

V. Results and Discussion

The best results from our experiments are summarized in Table II. The first row presents the performance of the dummy classifier, which classifies every sample into the majority class. As expected, the dummy classifier achieved the lowest accuracy and F1 scores, establishing a baseline for comparison. The first section of the table compares the performance of the Random Forest (RF) classifier applied to the data without any preprocessing. The raw data contained large embeddings, requiring RF to use fewer estimators due to the increased computational complexity. Notably, when dimensionality reduction through PCA was applied, the performance of the classifiers improved significantly. The second section of the table shows the results for each modality after PCA, where the number of features was reduced to the top 100 components.

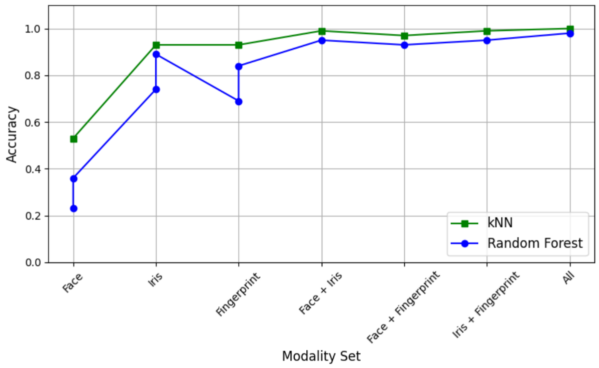

Fig. 2. Model performance based on modality combinations.

The subsequent sections highlight the performance of the classifiers when applied to multimodal data. Both RF and k-Nearest Neighbors (kNN) classifiers were evaluated on combinations of modalities, and we observed significant performance improvements when multiple modalities were fused. While RF is known for its robustness and generalization capabilities, kNN performed notably better for this task. The kNN classifier is particularly well-suited for this kind of problem because the embeddings represent data in a high-dimensional space, where instances from the same class are closer together. This spatial relationship is better leveraged by kNN than by RF, which does not inherently account for proximity between instances.

The combination of modalities provided the most notable improvements, particularly when the face modality was paired with other modalities like iris or fingerprint. Facial data alone performed relatively poorly, but when fused with iris or fingerprint data, the accuracy was very high, with the face + iris combination reaching an accuracy of 0.99 and an F1 score of 0.99. This highlights the complementary nature of the modalities and the significant impact of modality fusion.

As seen from Figure 2, kNN outperformed RF across all modality combinations, further confirming its suitability for this task. The highest performance was achieved with the full modality set (face + iris + fingerprint) using kNN with PCA, where both the accuracy and F1 score reached 1.00, reflecting perfect classification. These results underline the effectiveness of dimensionality reduction and modality fusion, with kNN emerging as the most suitable classifier for this biometric recognition task. In contrast, the RF classifier, while still effective, did not capitalize on the spatial relationships between data points as well as kNN, and thus performed slightly worse in multimodal settings. These findings demonstrate that multimodal fusion, particularly when combined with appropriate dimensionality reduction and classifiers like kNN, leads to substantial improvements in biometric recognition performance.

We also evaluated the performance of EmbraceNet on the same dataset. However, the model experienced overfitting, and the exit classifier was unable to perform better than the dummy classifier. This suggests that EmbraceNet did not generalize well on the data, likely due to the complexity of the task or insufficient model tuning. Further investigation into its hyperparameters and training process may be necessary to improve its performance on this biometric recognition task. However, given the excellent results achieved with late-fusion alone, introducing EmbraceNet would only add unnecessary complexity without significant gains.

VI. Conclusion

This work demonstrates the effectiveness of multimodal fusion for biometric recognition, combining iris, face, and fingerprint modalities to improve accuracy and robustness. By using late fusion techniques and efficient feature extraction models such as EfficientNet and FaceNet, we achieved significant performance improvements, particularly when combining multiple modalities. While advanced fusion methods like EmbraceNet showed promise, they were uneffective, mainly due to high complexity and overfitting. Our results highlight that simple late fusion, combined with appropriate dimensionality reduction and classifiers like kNN, can outperform more complex models, offering a balance of simplicity and effectiveness for practical biometric recognition systems. Future work may explore handling missing data and further investigating EmbraceNet's potential.

References

[1] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[2] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253511000558

[3] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[4] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *The Visual Computer*, vol. 38, no. 8, pp. 2939–2970, Aug 2022. [Online]. Available: https://doi.org/10.1007/s00371-021-02166-7

[5] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[6] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[7] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 07 2006. [Online]. Available: https://doi.org/10.1162/neco.2006.18.7.1527

[8] G. Bhatt, P. Jha, and B. Raman, "Representation learning using step-based deep multi-modal autoencoders," *Pattern Recognition*, vol. 95, pp. 12–23, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320319302146

[9] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[10] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[11] J.-H. Choi and J.-S. Lee, "Embracenet: A robust deep learning architecture for multimodal classification," *Information Fusion*, vol. 51, pp. 259–270, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517308242

[12] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[14] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[15] Biometric Ideal Test, "Biometric Ideal Test Database," http://biometrics.idealtest.org/#/, accessed: 2025-01-02.