IBB - Research Track Disposition: Multimodal biometric learning

Sebastijan Trojer

November 8, 2024

1 Topic Description

The goal of this topic is to compare recognition performance depending on what method for modality fusion is used. The experiment is divided into several steps:

1. Creating a new dataset from chosen modalities
2. Feature extraction on each modality (CNN based).
3. Modality fusion: Early/late fusion (baseline) and EmbraceNet fusion.
4. Result analysis.

1.1 Data

For this task we will use Casia-Iris-Thousand dataset for iris images, Casia-Fingerprint for fingerprint images and CelebFaces for face data. More information about the datasets is displayed in Table 1.

Dataset	# Subjects	Modality	# Samples
CelebFaces	10 177	Face	202 599
Casia-Iris-Thousand	1000	Iris	20 000
Casia Fingerprint v5	500	Fingerprint	20 000

Table 1: Modality Datasets

This selection of datasets gives 500 common subjects, which is sufficient for this task. Despite some of the datasets being outdated, the goal is not to reach SOTA performance but to determine best fusion techniques.

2 Related Work

Since handcrafted algorithms are mostly outdated we will focus on deep learning approaches for feature extraction, which will also streamline the process.

2.1 Modality Fusion

Modality fusion is the integration of different data types to enhance the performance of machine learning models. Unlike single-modality approaches, which rely solely on one type of input, modality fusion leverages the complementary strengths of different data sources.

Modality fusion can be categorized into two primary types [1]:

- Early Fusion: This approach combines raw data features from multiple modalities before model training. By integrating features at an early stage, the model learns joint representations that can capture the interplay between different data sources. Early fusion has shown success in applications such as audiovisual speech recognition, where synchronized feature extraction improves understanding of spoken words in noisy environments.

- Late Fusion: In contrast, late fusion involves training separate models for each modality and combining their outputs during the decision-making phase. This method offers flexibility, as each modality can be processed with a specialized model. Late fusion is particularly effective when data modalities have different characteristics or when models need to be fine-tuned independently.

There also exists something called hybrid or intermediate fusion, which consists of combining multi modal features of early and late fusion. A scheme illustrating the differences is shown on Figure 2.

In their survey on deep multimodal learning for computer vision, Bayoudh et al. [1] list a lot of other modality fusion approaches, some of them reaching as far back as to early 2000s, such as hidden Markov model based modality fusion [7], canonical correlation analysis based fusion [4], deep belief systems [5], step-based deep multi-modal audoencoders [2] and some more modern ones, like bimodal convolutional neural network based approach [6]. Peng et al. [8] investigated multimodal learning with transformers and provided a sophisticated survey on transformer based multimodal learning, providing resources on the topic.

For this task we decided to use EmbraceNet [3] for data fusion. EmbraceNet is a model consisting of two types of layers:

- Docking layers: Takes feature vectors of all modalities as input and converts them into dockable vectors, so that all vectors have the same shape.

- Embracement layer: Joins the vectors from the docking layers into a single vector based on multinomial sampling method.

Therefore the output is a single feature vector, that we pass to an exit classifier, as illustrated on Figure 1.

In the paper EmbraceNet fusion technique is also compared to multiple other approaches, such as early and late fusion, multimodal autoencoder and compact multi-linear pooling with some modalities or raw data missing, however EmbraceNet outperformed all of them.

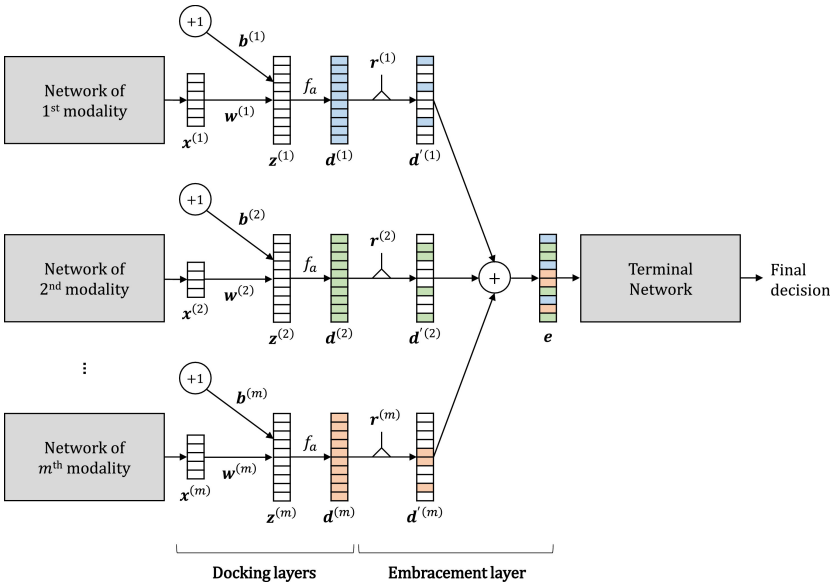

Figure 1: EmbraceNet structure [3].

References

[1] Khaled Bayoudh et al. "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets". In: *The Visual Computer* 38.8 (Aug. 2022), pp. 2939–2970. ISSN: 1432-2315. DOI: 10.1007/s00371-021-02166-7. URL: https://doi.org/10.1007/s00371-021-02166-7.

[2] Gaurav Bhatt, Piyush Jha, and Balasubramanian Raman. "Representation learning using step-based deep multi-modal autoencoders". In: *Pattern Recognition* 95 (2019), pp. 12–23. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2019.05.032. URL: https://www.sciencedirect.com/science/article/pii/S0031320319302146.

[3] Jun-Ho Choi and Jong-Seok Lee. "EmbraceNet: A robust deep learning architecture for multimodal classification". In: *Information Fusion* 51 (2019), pp. 259–270. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2019.02.010. URL: https://www.sciencedirect.com/science/article/pii/S1566253517308242.

[4] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. "Canonical Correlation Analysis: An Overview with Application to Learning Methods". In: *Neural Computation* 16.12 (2004), pp. 2639–2664. DOI: 10.1162/0899766042321814.

[5] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. "A Fast Learning Algorithm for Deep Belief Nets". In: *Neural Computation* 18.7 (July 2006), pp. 1527–1554. ISSN: 0899-7667. DOI: 10.1162/neco.2006.18.7.1527. eprint: https://direct.mit.edu/neco/article-pdf/18/7/1527/816558/neco.2006.18.7.1527.pdf. URL: https://doi.org/10.1162/neco.2006.18.7.1527.

[6] Lin Ma et al. "Multimodal Convolutional Neural Networks for Matching Image and Sentence". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015.

[7] L.R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286. DOI: 10.1109/5.18626.

[8] Peng Xu, Xiatian Zhu, and David A. Clifton. "Multimodal Learning With Transformers: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10 (2023), pp. 12113–12132. DOI: 10.1109/TPAMI.2023.3275156.

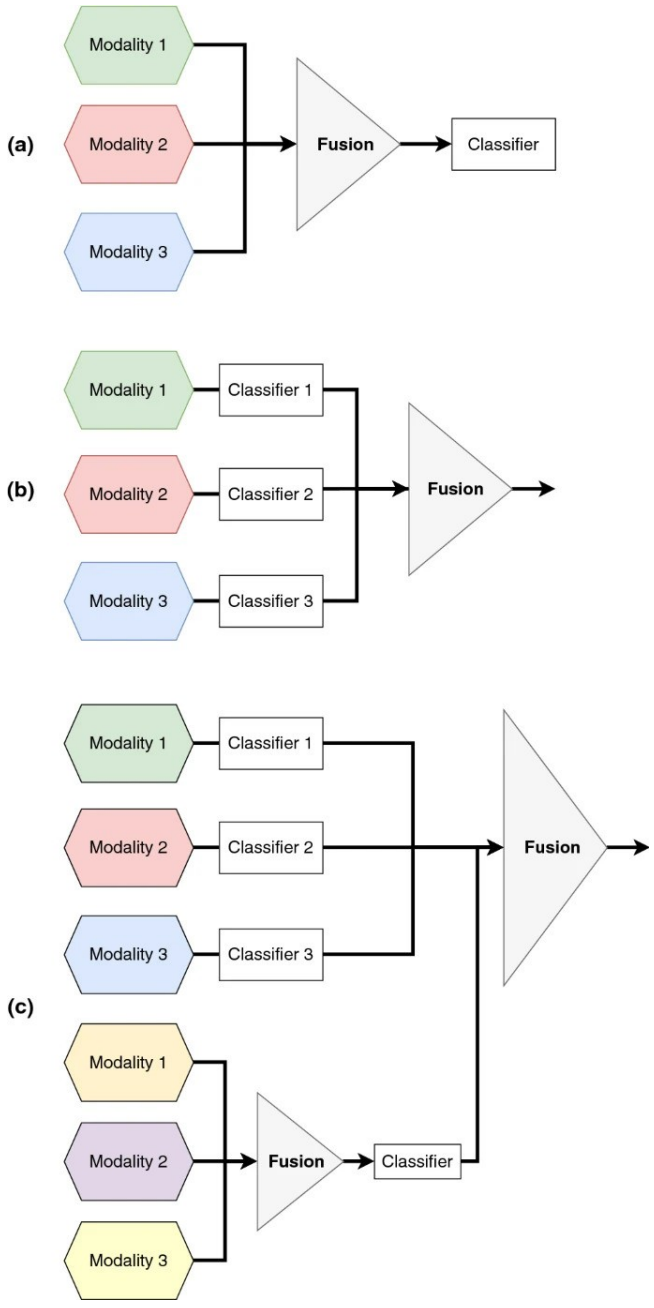

Figure 2: Early (a), late (b) and intermediate (c) fusion [1].