

Deep Combination of Radar with Optical Data for Gesture Recognition: Role of Attention in Fusion Architectures

Praveer Towakel, David Windridge and Huan X. Nguyen, *Senior Member, IEEE*

***Abstract*—Multimodal time series classification is an important aspect of human gesture recognition, in which limitations of individual sensors can be overcome by combining data from multiple modalities. In a deep learning pipeline, the attention mechanism further allows for a selective, contextual concentration on relevant features. However, while the standard attention mechanism is an effective tool when working with Natural Language Processing (NLP), it is not ideal when working with temporally- or spatially-sparse multi-modal data. In this paper, we present a novel attention mechanism, Multi-Modal Attention Preconditioning (MMAP). We first demonstrate that MMAP outperforms regular attention for the task of classification of modalities involving temporal and spatial sparsity and secondly investigate the impact of attention in the fusion of radar and optical data for gesture recognition via three specific modalities: dense spatiotemporal optical data, spatially sparse/temporally dense kinematic data, and sparse spatiotemporal radar data. We explore the effect of attention on early, intermediate, and late fusion architectures and compare eight different pipelines in terms of accuracy and their ability to preserve detection accuracy when modalities are missing. Results highlight fundamental differences between late and intermediate attention mechanisms in respect to the fusion of radar and optical data.**

***Index Terms*—Gesture recognition, deep learning, attention mechanism, multi-modality, radar combination.**

I. Introduction

A Multitude of applications have been found and refined over the years for gesture recognition including computer interactions, forensic identification, virtual environments, monitoring automobile drivers' alertness, medically monitoring patients and lie detection. To overcome the limitations of traditional recognition systems [1]–[3], researchers have turned to multimodal approaches that combine data from multiple sensors, thereby leveraging the strengths of multiple modalities to improve overall recognition performance.

Gesture recognition systems can be classified into two types, optical-based systems and non-optical-based systems. Recently there has been a lot of work done on both types. Optical-based gesture recognition studies mainly attempt to address the many challenges associated with using the camera which include illumination inconsistencies, motion blur and background clutter. A solution to this illumination problem has been proposed by [4] which is insensitive to environmental illumination and background variation by using a biologically inspired neuromorphic optical sensor that have microsecond temporal resolution, high dynamic range, and low latency. However, it is debatable if this technology will be widely available soon.

Recently the attention mechanism has emerged as a powerful tool for selectively focusing on relevant features and enabling more efficient and effective learning [5]. For selective attention, [6] used global visual descriptors and Common Spatial Patterns (CSP) to cleverly categorise videos taken by a humanoid robot. As for non-optical-based approaches such as radar or Wireless Fidelity (Wi-Fi) systems which do not suffer from the illumination problem, studies mainly focus on the accurate identification of gestures by recognising and profiling the details of hand gestures in different environments. [7] adopted the multisensor approach and presented a dual Doppler radar-based system which can capture subtle arm gestures with less positioning or environmental dependence. [8] proposed WiGRUNT, a WiFi-enabled gesture recognition system using a dual-attention network, to mimic how a keen human being intercepts a gesture regardless of the environment variations. While both approaches are making substantial progress, there remain several limitations, such as strict environmental requirements and low stability.

Radar is now a cheaper modality and a combination of optical and radar data by multimodal fusion has been seeing rapid development and integration [9]. This approach has the potential of combining the best of both modalities resulting in a system that is resistant to environmental changes while also performing adequately with low illumination. While fusion can solve many of the problems of gesture recognition, knowing when and how to fuse becomes an important consideration. Our approach focuses on finding the best fusion point for radar with optical and kinematic data and looking at the effects of having attention layers at different positions in fusion. We look at several metrics such as training time, accuracy and the ability of the system to retain active discriminant information while removing redundant information for different configurations of early, intermediate and late fusion. While we are cautious to generalise our results, we believe we obtained useful findings about the fusion of radar with optical and kinematic data in particular.

The main contributions of our work are as follows:

1) We present a new type of attention better suited for multi-modal sparse data applicable in both temporal and spatial dimensions. We show that this novel Multi-Modal Attention Preconditioning (MMAP) mechanism outperforms regular attention in accuracy, recall and f1-score when applied on sparse data, which could also offers broad potential beyond the specific application of gesture recognition investigated

Praveer Towakel is with Faculty of Science and Technology, Middlesex University (Mauritius campus). Email: Email: P.Towakel@mdx.ac.mu.

David Windridge is with the Department of Computer Science, Faculty of Science and Technology, Middlesex University, London, UK. E-mail: D.Windridge@mdx.ac.uk

Huan X. Nguyen is with the London Digital Twin Research Centre, Faculty of Science and Technology, Middlesex University London, UK, and also with International School, Vietnam National University, Hanoi, Vietnam. E-mail: H.Nguyen@mdx.ac.uk

Corresponding author: Huan X. Nguyen.

in this paper. Its utility is anchored in its specific handling of sparse data representations, a particular challenge when distributed across various modalities and domains. The mechanism's adaptive learning of cross-modal feature importances specific to the task at hand can effectively tackle data sparsity by focusing on informative signals amidst potential noise or null data points. This property has the potential to significantly improve performance in tasks involving varied sparse sensor readings, time-series analysis with missing data, natural language processing with sparse semantic embeddings, or any domain where data representations involve inherent intermittent modal sparsity.

2) Our approach leverages the potential of deep learning to unite multimodal features at an arbitrary level of abstraction, which we evaluate relative to early, intermediate and late fusion pipelines for 3 different modalities: radar, kinematic (skeleton) and optical data. We find that the multimodal fusion of different inputs in this context results in a clear improvement over unimodal approaches due to the complementary nature of the various input modalities. From the eight different architectures presented, it is noted that late fusion with late attention has the potential of outperforming early and intermediate fusion with all modalities present and also in circumstances where one of the modalities is masked.]
3) We present a comparative evaluation of early, intermediate and late fusion and indicate the point at which one particular modality can fail outright and another one can recover for the discrimination of classes. We also show that one modality can carry the discriminating information while other modalities fail. From our findings, end-to-end late fusion with pre-trained unimodal models can dynamically switch between different modalities based on their reliability.
4) We show that late fusion with late attention can recover information that is hidden in the fused decision softmax space. We found that late fusion using the attention mechanism can dynamically adapt the fusion strategy based on the context and highlight the most relevant information from each modality softmax output. From our findings, late attention recovered fully all information lost in the prediction of gesture classes. We found a jump in detection confidence of 40% with a Jaccard score of going from 0.53 to above 0.9.

II. Related Works

This section reviews in details the relevant work undertaken by other researchers that used radar, skeleton and optical data.

A. Visual features for gesture recognition

Before the advent of network-based approaches, gesture recognition was primarily based on computer vision techniques and handcrafted features. This involved identifying and extracting relevant features from images or videos of gestures, such as the shape, size, and movement of the hands, and then using classical machine learning algorithms, such as decision trees and support vector machines to classify the gestures based on these features. These systems were limited in their accuracy and ability to generalise to new and unseen gestures. For example, in [10] a global optimisation framework was made based on binary quadratic programming, [11] proposed a spatio-temporal feature named Mixed Features around Sparse key points, [12] proposed a scheme of aggregating the low-level polynormals into the super normal vector. However, due to the nature of these features, there is a risk that they fail to capture the relevant information from the input. Now, attention has shifted to network-based approaches where task-specific features can be obtained automatically. Most of the recent gesture recognition methods made use of deep learning such as [13]–[15] which boast superior results to isolated gesture recognition with handmade features. [16] proposed a system for dynamic hand gesture recognition using multiple deep learning architectures for hand segmentation, local and global feature representations, and sequence feature globalisation and the authors in [17] proposed two convolutional neural networks (CNNs) with partial and full weight sharing, for multimodal data where they employ both partial weight sharing and full weight sharing in such a way that modality-specific characteristics, as well as common characteristics across modalities, are learned from multimodal (or multi-sensor) data and are eventually aggregated in upper layers.

B. Other modalities for gesture recognition

As compared to wearable sensors, radar sensors are not affected by illumination and because of radar signal transmissivity, have the advantage of being robust even in the presence of occlusion. The works in [18]–[22] presented different systems which use radar for hand/body gesture detection where deep learning was proven to be excellent on analysing radar signatures for smart detection. In [23] 14 different hand gestures were studied and represented with signatures as a 3-dimensional tensor consisting of range-Doppler frame sequence. These signatures were then passed to a CNN to extract the unique features of each gesture. For the study, a low-power Ultra-wideband (UWB) impulse radar which transmits sharp temporal pulses was used. Recently, the research of gesture recognition on radar has opened a range of new possibilities in intelligent sensing. The radar sensor can instantaneously capture the range and speed of the gesture in each frame signal calculated by a fast Fourier transform (FFT). Radar technologies can be classified into two main categories: i) Pulsed radar [24], [25] and ii) Continuous-wave radar (CW) [18], [26], [27]. CW radars can further be classified as Frequency-Modulated CW (FMCW) radars and single-frequency CW (SFCW) radars. Several studies also denote SFCW radars as Doppler as they operate mostly on the Doppler effect. For classification, [19] used a two-antenna Doppler Radar to train a CNN to classify hand gestures where the beat signals from the two receiving antennas were used to generate feature arrays representing 14 different hand gestures. In the study 250 recorded samples for each of the 14 different gestures were used. This set was divided into 80% for the DCNN training with 5-fold validation and 20% for the testing. The training was repeated 5 times by rotating the testing set and the training set such that the average classification accuracy of training, validation and testing could be calculated. The testing results showed that the proposed two-antenna method outperformed previously used single-antenna approaches by 10% at a similar CNN level of complexity. In another study, [20] proposed a real-time gesture recognition system using a short-range radar, Soli, developed by Google. It was built from the

bottom up including signal processing, machine learning and communication. In the signal processing, 2D FFT was performed to generate the Range-Doppler Map (RDM) sequences in real time and clutters were removed using an adaptive background model based on Gaussian mixture model (GMM). The gesture was detected by the constant false alarm rate (CFAR) algorithm and then recognised by the long short-term memory (LSTM) encoder. The LSTM encoder extracted the global temporal features of the motion sequences. The study reported that the proposed system achieved high accuracy under various conditions. [28] proposed a novel two-step pipeline classification solution for surface-electromyography-based gesture (sEMG) recognition, which has been evaluated on 7 sparse multichannel and 4 high-density sEMG benchmark databases. The study also presented a cross-modal association model with adversarial learning to capture the intrinsic relationship between sEMG signals and hand poses. Experimental results indicated that compared with a crossmodal association model constructed without adversarial learning, the proposed model enables improved gesture recognition accuracy based on both sparse multichannel and high-density sEMG signals, although, the improvements achieved on sparse multichannel sEMG databases are stated to be higher than those achieved on high-density sEMG databases. In the framework analysis of [29], movement primitive, ie segmentation of a long sequence of human movement observation data is reviewed for its use to facilitate the identification of movement (this can also be applied to fingertip motions). The study proposed a framework to provide a structure and a systematic approach for designing and comparing different segmentation and identification algorithms. As can be seen from the variety and scope of the more recent papers, unimodal radar and optical studies are thriving through clever ways to compensate for the shortcomings of their respective modalities.

C. Multimodal fusion

Deep multimodal learning has achieved remarkable progress in recent years in multiple research areas [30], [31]. These methods fall under early, intermediate, or late multi-modal fusion. [30] presented a method for gesture detection and localisation based on multi-scale and multi-modal deep learning. In their method, spatial information was captured at a particular spatial scale (such as the motion of the upper body or a hand), and the whole system operated at two temporal scales. [31] described a novel method called Deep Dynamic Neural Networks (DDNN) for multimodal gesture recognition where a semi-supervised hierarchical dynamic framework based on a Hidden Markov Model (HMM) is proposed for simultaneous gesture segmentation and recognition where skeleton joint information, depth, and RGB images, are the multimodal input observations. Unlike most traditional approaches that rely on the construction of complex handcrafted features, their approach learns high-level spatiotemporal representations using deep neural networks suited to the input modality: a Gaussian-Bernouilli Deep Belief Network (DBN) to handle skeletal dynamics, and a 3D Convolutional Neural Network (3DCNN) to manage and fuse batches of depth and RGB images. [32] introduced a single-stage continuous gesture recognition framework, a Temporal Multi-Modal Fusion (TMMF), that can detect and classify multiple gestures in a video via a single model. This approach learns the natural transitions between gestures and non-gestures without the need for a pre-processing segmentation step to detect individual gestures. To achieve this, the authors introduced a multi-modal fusion mechanism to support the integration of important information that flows from multi-modal inputs, and which was scalable to any number of modes. Additionally, they propose Unimodal Feature Mapping (UFM) and Multi-modal Feature Mapping (MFM) models to map uni-modal features and the fused multi-modal features respectively. [33] proposed a method which first learns short-term spatiotemporal features of gestures through the 3-D convolutional neural network, and then learns long-term spatiotemporal features by convolutional LSTM networks based on the extracted short-term spatiotemporal features. In order to take full use of the advantages of 3DCNN and the attention mechanism, a combination of the two networks are utilised in the proposed deep architecture to learn spatiotemporal features.

III. RADAR DATA CAPTURE

In this work, we investigate the fusion of radar with other modalities where radar data capture dictates the type and density of radar frames. The experiment in this study uses the evaluation version of AWR1642 which consists of a waveform generator, an antenna array with two transmitters and four receivers, a signal de-modulator and an analogue-to-digital converter (ADC) converter. The waveform generator transmits the chirp signal through the transmit antenna. Then an intermediate frequency (IF) signal is obtained using a low-frequency filter (LPF). The radar model consists of x-y positions, intensity and range data. 1D FFT processing is performed for range while 2D (velocity) FFT processing produces the velocity. CFAR detection in range direction uses the mmWave library.

A. Range measurement

In a radar system, the transmitter emits a high-frequency electromagnetic signal that is directed towards the object. As the signal reflects off the object, it returns to the receiver. The receiver then analyses the reflected signal and measures the Doppler shift to determine the object's distance. The time delay (τ) can be obtained from the distance d to the detected object, $\tau = 2d/c$, where c is the speed of light.

B. Velocity measurement

To measure the velocity of multiple points of the hand/arm a radar system must transmit multiple chirps. This can be a set of N equally spaced chirps. For the case of two point, we first compute the range-FFT of the reflected set of chirps, and get a set of N identically located peaks but with different phases which incorporate the individual contributions from each of the points. To this, a second FFT, called Doppler-FFT, can be performed on the N phasors to resolve the points. The velocity measurement is carried out as

$$v = \frac{\lambda\omega}{4\pi T_c} \tag{1}$$

where ω is the discrete frequency corresponding to the phase difference between consecutive chirps, λ is the wavelength and T_c is the chirp period.

C. Angle estimation

Angle estimation exploits a similar concept. Angle estimation requires at least 2 RX antennas. What is exploited here is the differential distance of the object to each of these antennas. So the transmit antenna transmits a signal that is a chirp. It is reflected off the object, and one ray goes from the object to the first RX antenna and another ray goes from the object to the second RX antenna. This results in a phase change in the peak of the range-FFT or Doppler-FFT. This result is used to perform angular estimation, using at least two RX antennas. The differential distance from the object to each of the antennas results in a phase change in the FFT peak. The phase change enables the estimation of the angle of arrival (AoA). Under the assumption of a planar wavefront basic geometry shows that $\Delta d = I \sin(\theta)$, where I is the distance between the antennas. Thus the angle of arrival (θ), can be computed from the measured phase change $\Delta\Phi$:

$$\theta = \sin^{-1}\left(\frac{\lambda\Delta\Phi}{2\pi I}\right) \tag{2}$$

Note that $\Delta\Phi$ depends on $\sin(\theta)$. This is called a nonlinear dependency; $\sin(\theta)$ is approximated with a linear function only when θ has a small value: $\sin(\theta) \sim \theta$.

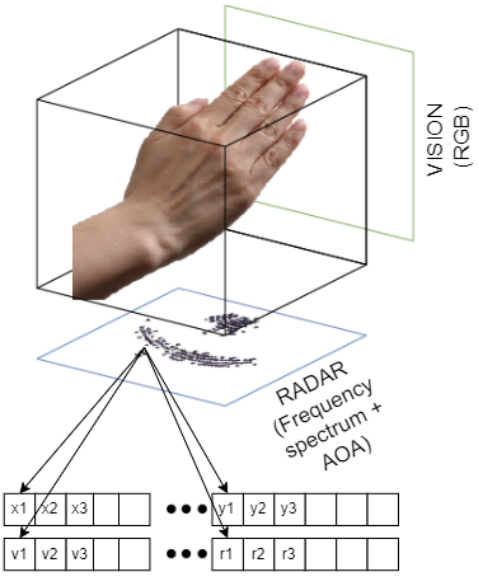

Fig. 1: Shows the data captured by radar, ranges(r), x-positions, y-positions and velocities(v).

The relationship between radar and vision is such that radar data can be viewed as a sparsely populated, low-resolution image that contains valuable depth and velocity information (as shown in Fig. 1). This is in contrast to high-resolution optical data or continuous kinematic data. As a result, when fusing radar data with other modalities, the complementary nature of the information it provides must be considered to avoid introducing noise into the system. To effectively utilise the different modalities and create a universal decision maker, contextualised attention is needed. This approach takes into account the context of the data and their interrelationships, allowing for more effective integration and decision-making. By carefully weighing the strengths and weaknesses of each modality, contextualised attention can help to identify the most appropriate use of the data to achieve the desired outcomes. Ultimately, the goal is to develop a system that can leverage the strengths of all modalities to make informed decisions and improve overall performance.

IV. Multimodal Gesture Recognition with Attention

In this section, we present architectures for learning multimodal time series data (optical, skeletal and radar) for class prediction. First, we introduce the attention mechanism for multimodal time series data and subsequently describe early, intermediate and late fusion architectures as shown in Fig. 2, Fig. 3 and Fig. 4.

A. Attention

The attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, key, value and output are all vectors [5]. The keys and queries are dotted and multiplied to obtain the corresponding attention weights, and finally, the obtained weights and values are dotted to obtain the final output. For self-attention, the three matrices Q (Query), K (Key), and V (Value) are all from the same input. We first compute the dot product between Q and K and then divide the result by a scale $\sqrt{d_k}$ to prevent the result from being too large, where d_k is the dimensionality of a query and key vector. The result is then normalised using the Softmax operation and then multiplied by the matrix V to obtain the representation of the weight summation. This computational procedure can be expressed as follows.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

B. Proposed Multi-Modal Attention Preconditioning

This proposed MMAP function, whilst also seeking to map a query and a set of key-value pairs to an output, however, incorporates an additional trainable attention weight matrix and a dropout mechanism which are not present in regular attention mechanism. This helps to emphasise relevant features and minimise the effect of non-informative or noisy features, which is especially beneficial for handling sparse multi-modal data such as radar readings. This can be represented as follows: given a set of queries Q, keys K, and values V, MMAP first applies the learned linear transformations:

$$Q' = W_q \cdot Q$$

$$K' = W_k \cdot K$$

$$V' = W_v \cdot V, \tag{4}$$

where W_q, W_k and W_v are the learned weight matrices of the corresponding transformation functions. We then compute the dot product between the transformed queries and keys divided by a scale factor $\sqrt{d_k}$ to prevent divergence, with d_k the dimensionality of a query and key vector. This result is then passed through a softmax function to normalise the scores, creating the attention weights A, which are also subjected to dropout mechanism:

$$A = \text{dropout}\left(\text{softmax}\left(\frac{Q'K'^T}{\sqrt{d_k}}\right)\right). \tag{5}$$

The output of the function is then the weighted sum of the values, passed through a final learned linear transformation:

$$\text{AttentionPreconditioning}(Q, K, V) = W_o \cdot (A \cdot V') \quad (6)$$

where W_o is the learned weight matrix of the output transformation. A pseudocode description of this process is given in Algorithm 1.

Algorithm 1 Multi-Modal Attention Preconditioning (MMAP)

procedure ATTENTIONPRECONDITIONING(Q, K, V, H, N, S, D)
  Initialise transformation matrices $W_q, W_k, W_v, W_o \in \mathbb{R}^{H \times H}$
  $Q' = Q \times W_q$, $K' = K \times W_k$, $V' = V \times W_v$
  Reshape Q', K', V' into Q'', K'', V'' with N heads
  Compute attention scores $S = \frac{Q'' \times (K'')^T}{\sqrt{S}}$
  Compute attention weights $A = \text{softmax}(S)$, apply dropout $A = \text{dropout}(A, D)$
  Compute attended output $O' = A \times V''$
  Reshape O' and apply final linear transformation $O = \text{reshape}(O') \times W_o$
  return O
end procedure

In multimodal fusion, intermediate attention can be used to selectively attend to the most informative parts of the intermediate feature representations, before they are combined using a fusion method such as concatenation or averaging. This can help recover any "hidden" information that may have been lost during the individual processing of each modality. Attention can also be used to selectively attend to the most informative parts of the decision softmax layer in the fusion process and in such cases is called late attention.

C. Architectures

Multimodal fusion is a method that integrates different modalities with different properties. The fusion of information from different modalities is a common approach to improving performance. The essence is to combine heterogeneous sensor data to enable the implementation of complementary information processes. The networks in this work contain several elements, which are described as follows:

- Input: Time series data are appended one after another to make a sequence. Input blocks made are in the form of (a, s, f) where a are the actions performed s are the sequences and f are the features.
- Convolution layer: This layer creates a convolution kernel that is convolved with the layer input to produce a tensor of outputs.
- Concatenation layer: It takes input as a list of tensors, all of the same shape except for the concatenation axis, and returns a single tensor that is the concatenation of all inputs.
- Addition layer: It takes as input a list of tensors, all of the same shape, and returns a single tensor that is the addition of all tensors (also of the same shape).

1) Early fusion (EF): Fig. 2 shows our architecture for early fusion with and without attention. Early fusion consists of integrating the separate raw data modalities into a unified representation before proceeding through the learning/feature extraction process.

2) Intermediate fusion (IF): Fig. 3 shows our architecture for intermediate fusion with and without attention. Intermediate fusion combines the features that distinguish each type of data to produce a new representation that is more expressive than the separate representations from which it arose. For example, the fusion of features extracted from images and those extracted from skeletal sequences, allows us to take advantage of the strengths of both representations simultaneously (such as resolution from optical data and depth information from radar data). We denote the single model as $h(\cdot)$. The final prediction can be written as

$$p = h\left([\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m]\right) \quad (7)$$

where $\boldsymbol{v}_i, i \in \{1, 2, \cdots, m\}$ is the ith element of m modalities.

3) End-to-end late fusion (LF): Fig. 4 shows our architecture for end-to-end late fusion with attention where the score of merging is computed by a deep neural network. In particular, for m considered modalities, we used pre-trained architectures to generate score vectors from each modality individually. Each such architecture performs both feature extraction and classification and provides a vector of the potential membership scores to each of the considered classes. After being pre-processed, these vectors are used as inputs to our network for training. Such an operation allows us to learn more consistent joint decisions than conventional merging rules.

D. Sparsity

Spatial sparsity is used in our work to describe the number of pixels of an image or elements in a vector that are not populated. Highly populated images/vectors have low sparsity while images/vectors with low pixels/elements are highly sparse. Temporal sparsity in this work refers to temporal population of images/vectors.

V. Experiment, results and discussion

In this section, we present the data capture method and experimental settings used in our experiments.

A. Experimental settings

All models are implemented on Tesla a P100 16GB GPU under Linux environment. The deep learning models are optimised by mini-batch gradient descent with the Adam [35] optimiser and a maximum number of epochs of 200. We select the hyperparameters which have obtained a minimum loss on the validation set. To the best of our knowledge, there is no existing dataset with optical, radar and kinematic data for gesture recognition. For the purpose of testing our early, intermediate and late pipelines, data was collected for 12 gestures. This data is in the format of 30 frames per sample for 20 samples per action for 12 actions. This was collected for optical and radar data. The entire dataset has been made available at [36] in a NumPy format. We also want to take advantage of skeletal information and there exists a very effective framework for the extraction of kinematic skeleton data for gestures, Google's MediaPipe framework. MediaPipe [37] is

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3307768

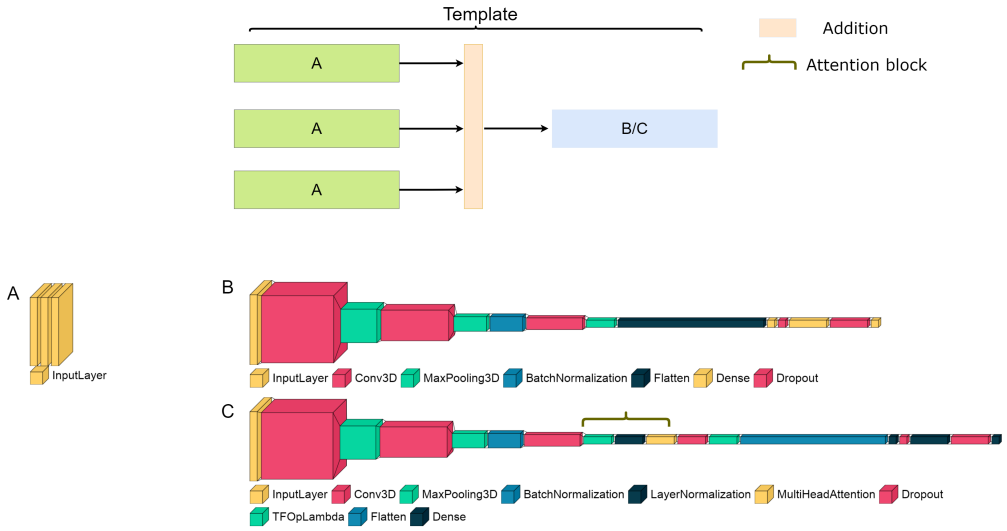

Fig. 2: Modular pipeline combination: (A/B) early fusion, (A/C) early fusion with intermediate attention. [34]

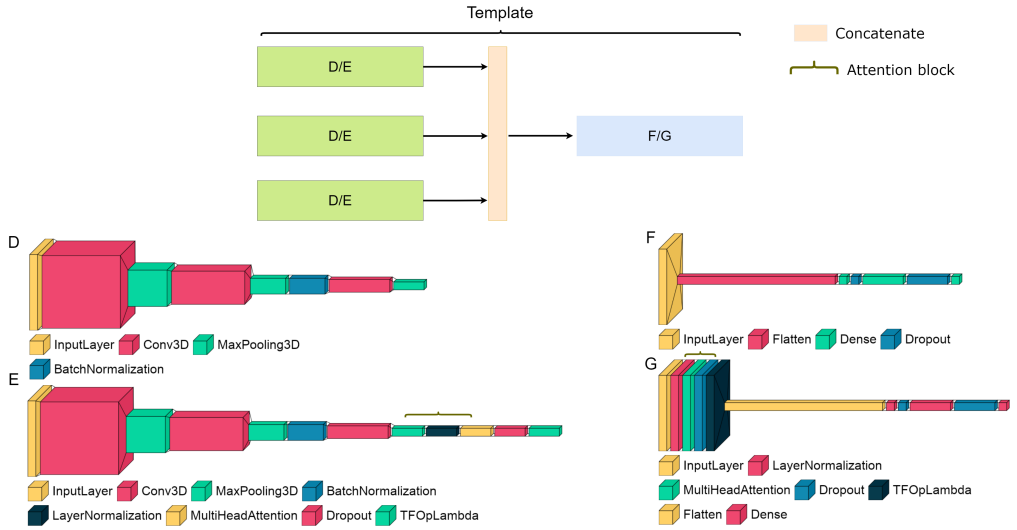

Fig. 3: Modular pipeline combination: (D/F) intermediate fusion, (E/F) intermediate fusion with intermediate attention, (D/G) intermediate fusion with late attention.

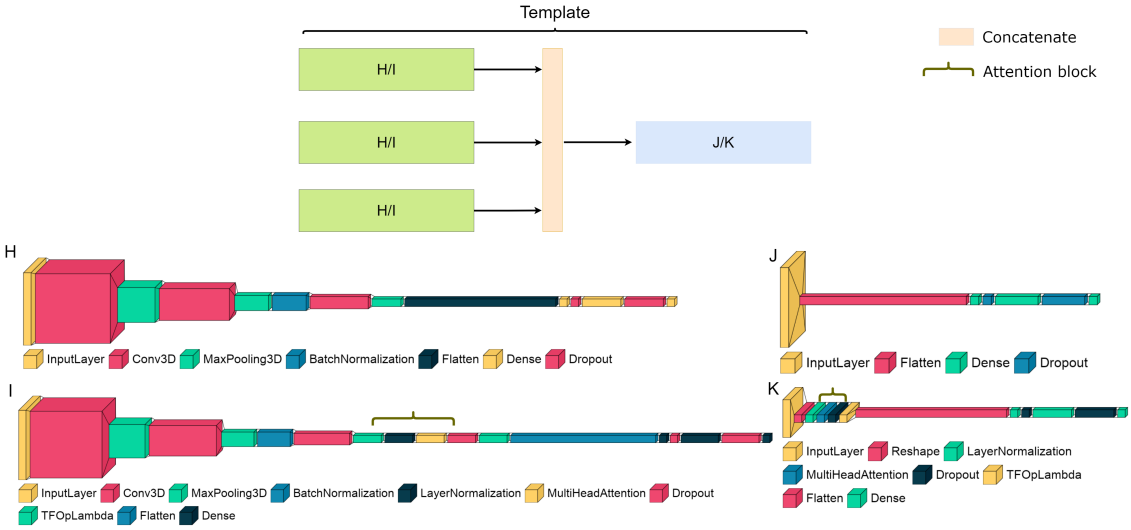

Fig. 4: Modular pipeline combination : (H/J) late fusion, (I/J) end-to-end late fusion with intermediate attention (pre-trained unimodal models), (H/K) late fusion with late attention.

an open-source framework, used for media processing. It is a framework for building machine learning pipelines for processing time-series data. This hand-tracking solution utilises a machine-learning pipeline consisting of two models working together. A palm detector that operates on a full input image and locates palms via an oriented hand bounding box and a hand landmark model that operates on the cropped hand bounding box provided by the palm detector and returns high-fidelity 2.5D landmarks. The MediaPipe pipeline returns 63 points corresponding to 21 joints for each hand. We compared the performance of different models with the following metrics derived from the confusion matrices.

1) Overall Accuracy (OA): This metric represents the proportion of correctly classified samples in all tested samples, and is computed by dividing the number of correctly classified samples by the total number of test samples.
2) Loss: This is a value that represents the summation of errors made by the model during the learning process.
3) Recall: Also known as sensitivity or true positive rate (TPR), this metric represents the proportion of actual positives that are correctly identified.
4) F1 Score: The F1 score is the harmonic mean of precision and recall, taking both metrics into account. It is particularly useful in the case of uneven class distribution, as the F1 score values the balance between precision and recall.

To measure performance we use the Jaccard index. This relies on frame-by-frame prediction accuracy. More precisely, if GT_i denotes the sequence of ground truth labels in video i, and R_i the algorithm output, the Jaccard index score of the video is defined as:

$$JS_i\left(GT_i, R_i, g\right) = \frac{N_s\left(GT_i, R_i, g\right)}{N_u\left(GT_i, R_i, g\right)}, \tag{8}$$

$$\text{and } JS_i = \frac{1}{\mid \mathcal{G}_i \mid} \sum_{g \in \mathcal{G}_i} JS_i\left(GT_i, R_i, g\right), \tag{9}$$

where $N_s\left(GT_i, R_i, g\right)$ denotes the number of frames where the ground truth and the prediction agree on the gesture class g. The quantity $N_u\left(GT_i, R_i, g\right)$ reflects the number of frames labelled as a gesture g by either the ground truth or the prediction, and $\mathcal{G}_i$ denotes the set of gestures either in the ground truth or detected by the algorithm in the sequence i. The average of the JS_i over all test videos is reported as the final performance measure. We have conducted the experiments and drawn some interesting observations from the results, which we summarise in the remarks followed up by explanations.

B. Training results

1) Unimodal models: For reference, we trained six unimodal models: optical, radar and skeletal data with and without the attention mechanism.

Remark 1: *Attention is an effective mechanism to address the sparse and resolution-poor nature of the millimeter wave radar modality. Our experiments show an increase of around 6% in overall accuracy and a decrease in training time of 8%.*

For radar, the convolution layer retains relative position and thus attention followed by application of softmax per feature map attenuates the Doppler component attributable to interference and amplifies the Doppler component dattributable to valid gestures. The attention mechanism learns to distinguish valid Doppler components from invalid ones effectively via the intrinsic pattern of the Doppler signal. As for skeletal data and optical data, an architecture with the attention mechanism UM-O-IA and UM-M-IA outperformed their basic variants in training time, with a decrease of 8% (Table I). We also note lower loss values for the optical data.

TABLE I: Performance metrics of unimodal model (UM) for optical/mediapipe(skeletal)/radar data (O/M/R) with or without intermediate attention (IA)

Architecture	Val. Accuracy	Val. Loss	Val. Recall
UM-O	1.0	0.033	1.0
UM-O-IA	1.0	0.013	1.0
UM-M	0.958	0.105	0.958
UM-M-IA	0.972	0.086	0.972
UM-R	0.889	0.404	0.875
UM-R-IA	0.944	0.344	0.931

From the results of validation accuracy, loss and recall (Table I), we can draw the following insight:

Remark 2: *Radar data gives the worst performance considered in isolation.*

While architectural biases can be at play for small variations in accuracy, we suggest that this large discrepancy for radar is a direct consequence of the sparse and resolution-poor nature of the radar modality. Also, the original echo of the hand gesture may contain some random interferences, which may mislead the training of the neural network classifier and reduce the speed of convergence and recognition rate.

2) MMAP performance on dense and sparse data: When comparing MMAP with regular MHA, implemented in the context of a transformer model for high sparsity tasks, our goal was to investigate the impacts of these attention mechanisms on model performance in terms of accuracy, F1 score and recall. These experiments were conducted in a cross-validation setting to ensure robust and unbiased comparisons.

Remark 3: Attention mechanism efficacy is data-dependent.

As shown in the results (Table II), standard MHA worked best for optical (average validation accuracy of 1), an inherently dense data type. In contrast, MMAP performed better on the intrinsically sparse data types: Skeletal (average validation accuracy of 1) and Radar (average validation accuracy of 0.92).

Remark 4: Specialised attention mechanisms can surpass more general ones.

MMAP consistently performed better on sparse data types. Skeletal (S) and Radar (R) models using MMAP outperformed their MHA counterparts, validating the use of specialised mechanisms, such as MMAP with its additional attention weight matrix and dropout mechanism, for specific modality data types.

Remark 5: Temporal density can compensate for spatial sparsity and MMAP can capitalise on this feature.

Models trained on Skeletal (S), which is temporally dense but spatially sparse, achieved high performance (validation accuracy of 1 for MMAP and 0.98 for MHA), suggesting the relative importance of temporal density. This result shows that MMAP can effectively utilise temporal information to make accurate predictions.

Remark 6: MMAP demonstrates substantial performance improvement for radar data

The significant performance increase for Radar (R) data with MMAP (validation accuracy increase from 0.81 to 0.92) empha-

sises the promise of MMAP in handling sparse data types. This is likely because MMAP's specialised mechanism, designed to minimise the effect of non-informative features, suits well for sparse data types like radar.

TABLE II: Comparison of validation metrics between Multi-modal Attention Preconditioning (MMAP) and Multi-head Attention (MHA) for different modalities

Modality	Attention	Val Loss	Val Accuracy	Val Recall	Val F1
S	MMAP	0.000442	1	1	1
S	MHA	0.114136	0.979167	0.984375	0.984375
O	MMAP	0.047544	0.979167	0.984375	0.984375
O	MHA	0.022415	1	1	1
R	MMAP	0.372836	0.916667	0.90625	0.90625
R	MHA	0.905177	0.8125	0.854167	0.854167

3) Fusion models: The resulting training characteristics of radar fusion with optical and skeletal using different configurations for EF (Fig. 2), IF (Fig. 3 and Fig. 4) are shown in 8 different architecture in Table III. They are early fusion (EF), early fusion with attention (EF-IA), intermediate fusion (IF), intermediate fusion with intermediate attention (IF-IA), intermediate fusion with late attention (IF-LA), late fusion (LF), end-to-end late fusion with intermediate attention (LF-IA) (pre-trained unimodal models), end-to-end late fusion with late attention (LF-LA) (pre-trained unimodal models). From Table III, late attention produced

TABLE III: Performance of 8 different architectures.

Architecture	Val. Accuracy	Val. Loss	Val. Recall
EF	1.0	0.0012	1.0
EF-IA	0.986	0.1045	1.0
IF	1.0	0.0006	1.0
LF	1.0	0.5548	0.875
IF-IA	1.0	0.001	1.0
LF-IA	0.986	0.9197	0.375
IF-LA	1.0	0.0003	1.0
LF-LA	1.0	0.0005	1.0

the best results for intermediate and late fusion with validation loss for LF-LA being 0.0005 and 0.0003 for IF-LA being the lowest loss recorded.

C. Real-time ablation tests

To find the characteristics of different fusion strategies we set up two pipelines (shown in Fig. 3 and 4). The intermediate and late pipelines are trained using the rectified Adam optimiser with a learning rate of 0.005 for 200 epochs with an early stopping of 20 patience by using mini-batches of 64 samples. We choose the final optimal models based on the performance on the validation set. The end-to-end late fusion pipelines were constructed with pre-trained unimodal models concatenated at the decision layer followed by dense and softmax layers. The models were then trained on the OMR dataset [36] and achieved their best performance for multi-label classification at self-attention block with sixteen heads with $d_k = d_v = 128$, where d_k, d_v are the dimension of the key and value respectively, in all the self-attention blocks. To compare the effectiveness of fusion methods with spatiotemporally sparse data (radar), spatially sparse and temporally dense data (skeleton), and dense spatiotemporal data (optical), experiments with real-time data streams were set up where the effects of each modality are investigated separately in different test runs. The experiment tests both pipelines on the 12 gestures with a video feed producing optical/MediaPipe (skeleton) streams of data with the shape (-,48,64,3) and (-,1,126) respectively per second stacked into 30 frames arrays. Radar data was collected in a data stream of shape (-,1,1000) and then converted to radar frames. To best show the results from our real-time test we used the Pearson-moment correlation (PPM) to compare test cases. Fig. 6 represent the detection confidence for each of the 12 gestures with all modalities. The models return real-time detection of the most recent 30 frames window.

Remark 7: *The point at which fusion occurs dictates the characteristics of the resulting model.*

Remark 7 is supported by Fig. 6 which shows that even though all test architectures scored close to 100% (table III) on testing with fully featured 30 frames pre-recorded data there is a wide variety of profiles produced when a real-time 30 frames window is used with the current lastest frame for detection confidence.

Remark 8: *Late attention can significantly improve detection accuracy for late fusion*

Remark 8 is supported by the difference between LF-LA, LF-IA and LF in Fig. 6 (and respective Jaccard index in table IV) where an LF-LA produced a Jaccard index of 0.916 which is better than either LF-IA with 0.380 and LF with 0.533. This test shows higher confidence level of late fusion with late attention.

Remark 9: *LF-LA produced high confidence from the first frame of detection, which is much faster than any IF pipelines.*

Remark 9 is supported by Fig. 6 which shows LF-LA detection "snapping" from the detection of one gesture to another within 1-2 frames. The IF counterparts are noted to do this within 4-5 frames. When all modalities are present, LF-LA outperforms the IF variants on detection speed. We note that from the correlation matrices, the IF models produced fewer instances of false positives. LF and LF-IA performed the worst when all modalities are present on detection confidence.

TABLE IV: Jaccard index with all modalities (G: Gesture)

G	LF-IA	EF	IF-LA	LF	LF-LA	IF-IA	IF	EF-IA
1	0.274	1.0	1.0	0.558	1.0	1.0	1.0	1.0
2	0.247	0.944	0.929	0.456	0.964	0.931	0.988	0.954
3	0.218	0.8	0.759	0.454	0.899	0.766	0.928	0.865
4	0.479	0.973	0.751	0.579	0.921	0.964	0.922	0.966
5	0.428	0.95	0.957	0.622	0.932	0.929	0.949	0.721
6	0.46	0.89	0.809	0.537	0.866	0.948	0.932	0.785
7	0.598	0.888	0.694	0.507	0.879	0.556	0.843	0.898
8	0.511	0.912	0.619	0.635	0.85	0.965	0.897	0.832
9	0.313	0.861	0.915	0.568	0.883	0.652	0.838	0.791
10	0.385	0.903	0.925	0.424	0.944	0.619	0.787	0.817
11	0.354	0.94	0.893	0.487	0.893	0.921	0.86	0.876
12	0.298	0.702	0.951	0.563	0.956	0.931	0.906	0.91

From Table IV, the architecture with the highest Jaccard index across all 12 gestures is EF, with values ranging from 0.702 to 1.0. The architecture with the lowest Jaccard index is LF-IA, with values ranging from 0.247 to 0.598. The other architectures have Jaccard indices ranging from 0.313 to 0.635, indicating moderate to high similarity between the predicted and ground truth segmentations.

The data in Table V shows that different architectures have varying performances, with some architectures performing better than others. The highest overall Jaccard index is 0.916, which is achieved by the LF-LA architecture. The lowest overall Jaccard index is 0.38, which is achieved by the LF-IA architecture. The other architectures have Jaccard indices between 0.533 and 0.904.

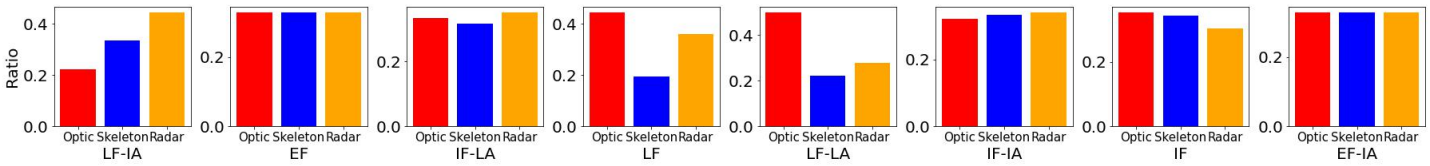

Fig. 5: Shows bias to optical, skeletal and radar data

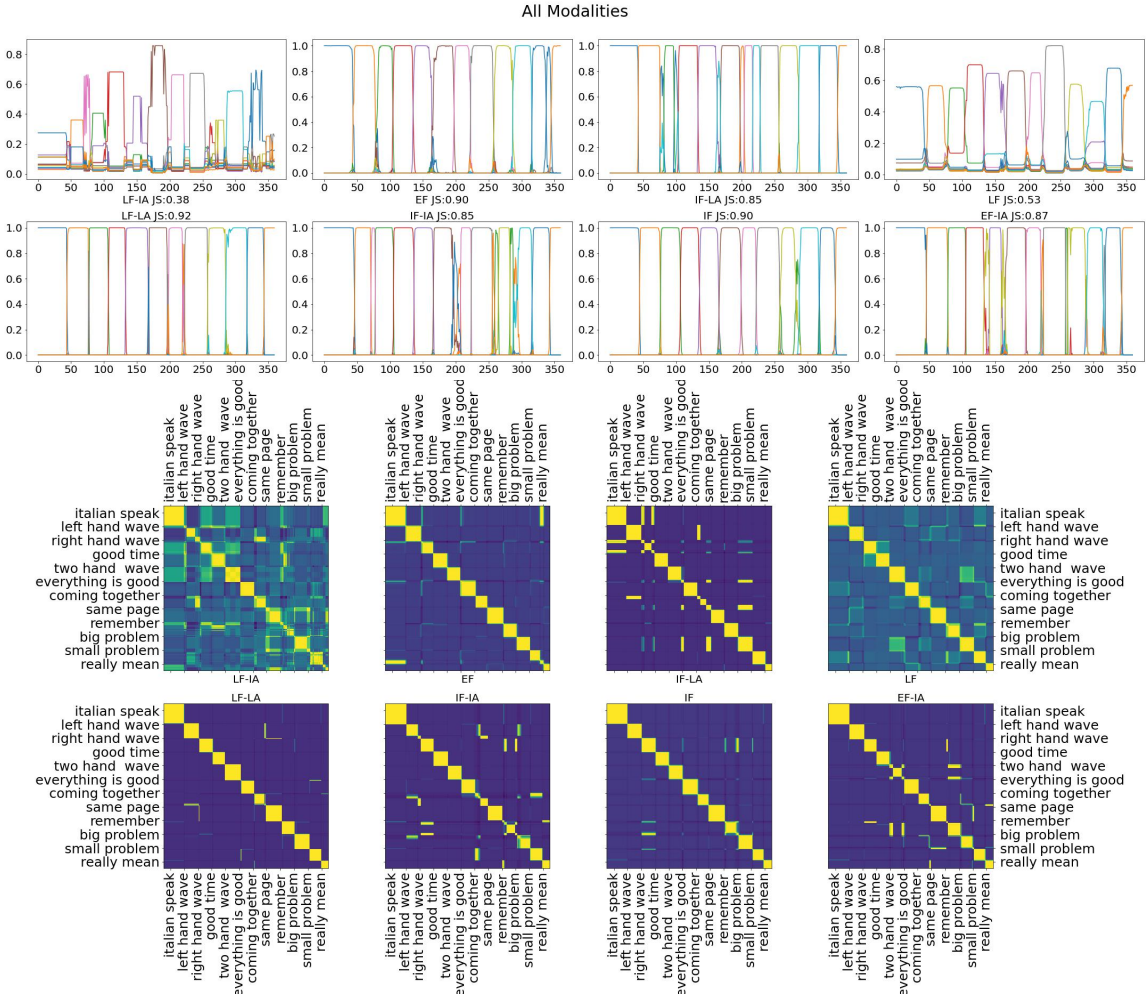

Fig. 6: Performances of the 8 different architectures, EF: early fusion, EF-IA: early fusion with attention IF: intermediate fusion, IF-IA: intermediate fusion with intermediate attention, IF-LA: intermediate fusion with late attention, LF: late fusion, LF-IA: end-to-end late fusion with intermediate attention (pre-trained unimodal models), LF-LA: end-to-end late fusion with late attention (pre-trained unimodal models). (b) Pearson-moment correlation (PPM) to compare test cases.

TABLE V: Overall Jaccard index all modalities

Jaccard index	Architecture	**Jaccard index**	Architecture
0.38	LF-IA	0.916	LF-LA
0.897	EF	0.848	IF-IA
0.85	IF-LA	0.904	IF
0.533	LF	0.868	EF-IA

1) Data masking: In co-learning, all the modalities are present at training time and some are missing at test time. The modalities which are not present at test time can be supported by other modalities during training. To investigate the effects of full modality removal, tests were performed with a setup having the optical, skeletal, and radar components completely removed successively. The presented PPM diagrams (Fig. 7,8,9) consist of a PPM of our reference all modality results (lower triangle) juxtaposed with the relevant PPM under investigation (upper triangle).

a) Optical data masking: With the optical modality masked, we can observe a substantial deterioration of IF with a Jaccard index going from 0.904 to 0.662 (profile shown in the PPM in Fig. 7). LF shows significant deterioration with a Jaccard index going from 0.533 to 0.242. This points to IF and LF performing poorly when the optical modality is absent. LF-LA on the other hand suffered degradation going from 0.916 to 0.692 but retained clear and distinct classes. This points to better preservation of both the complementarity and redundancy of the different modalities.

Remark 10: *IF-IA and IF-LA (Fig. 7) produced a lower Jaccard*

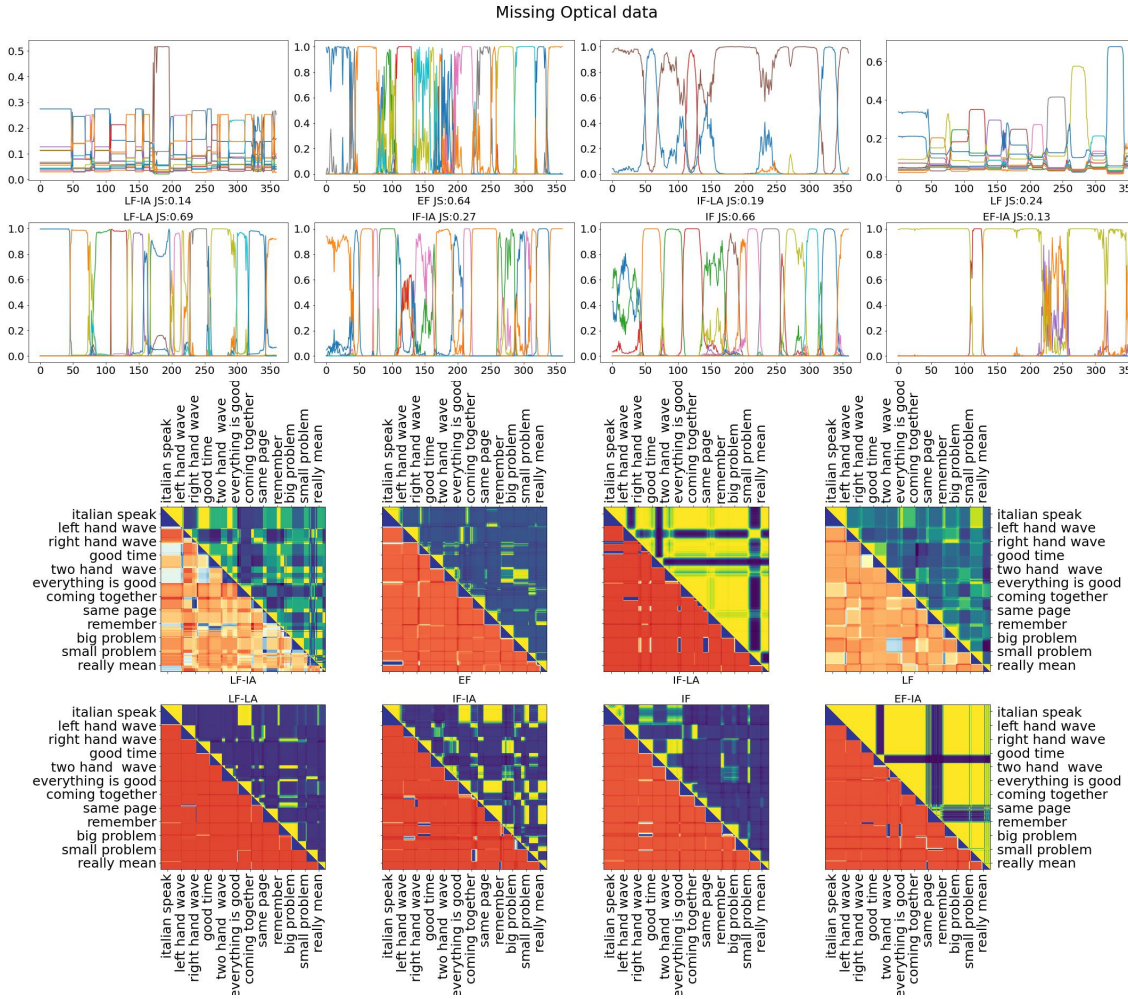

Fig. 7: PPM diagrams consist of a PPM of our reference all modality results (lower triangle) juxtaposed with the relevant PPM under investigation (upper triangle) for optical data missing

index score than the variants with no attention.

This is an interesting behaviour as it points to the attention mechanism at an intermediate stage introducing noise and thus making the models more susceptible to changes from the input. This shows that attention is not always desirable and can produce worst results than its basic variant when modalities are masked for **intermediate fusion**. To note the distribution (Fig. 5) shows that 30% of the neurons in the network were found to carry skeletal and radar information.

TABLE VI: Jaccard index with optical mask

G	LF-IA	EF	IF-LA	LF	LF-LA	IF-IA	IF	EF-IA
1	0.274	0.95	0.0	0.0	0.997	0.0	0.442	0.0
2	0.137	0.748	0.0	0.079	0.952	0.303	0.57	0.0
3	0.0	0.541	0.0	0.152	0.716	0.726	0.87	0.0
4	0.092	0.727	0.209	0.262	0.881	0.537	0.883	0.302
5	0.078	0.406	0.262	0.291	0.723	0.157	0.491	0.291
6	0.139	0.0	0.499	0.187	0.242	0.0	0.161	0.0
7	0.316	0.417	0.5	0.202	0.19	0.03	0.535	0.0
8	0.188	0.791	0.0	0.279	0.523	0.36	0.945	0.0
9	0.115	0.707	0.0	0.364	0.85	0.0	0.734	0.397
10	0.097	0.914	0.0	0.33	0.514	0.015	0.627	0.5
11	0.119	0.852	0.391	0.39	0.861	0.323	0.814	0.0
12	0.138	0.656	0.399	0.373	0.852	0.821	0.867	0.024

Table VI shows the Jaccard index values when the optical modality is removed. The removal of the optical modality has a significant impact on the performance of the different architectures, with several of them showing Jaccard index values close to 0.0. On the other hand, some architectures still show relatively good performance even when the optical modality is removed, such as LF-LA (0.997) and IF (0.945).

TABLE VII: Overall Jaccard index optical mask

Jaccard index	Architecture	**Jaccard index**	Architecture
0.141	LF-IA	0.273	IF-IA
0.642	EF	0.662	IF
0.188	IF-LA	0.126	EF-IA
0.242	LF	0.692	LF-LA

Table VII shows the overall Jaccard index for different architectures with the optical modality removed. The highest score is for the architecture "LF-LA" with a Jaccard index of 0.692. The lowest score is for the architecture "EF-IA" with a Jaccard index of 0.126. The other architectures have scores in between these two extremes.

b) Skeletal data masking: When the skeletal modality was masked, LF-LA produced the highest Jaccard index of 0.745, the closest second was the degraded EF with 0.686. We note that in this circumstance the architectures that performed the worst are IF-LA with 0.180 and LF-IA with 0.170. LF-IA produced a higher number of false positives as compared to IF architectures

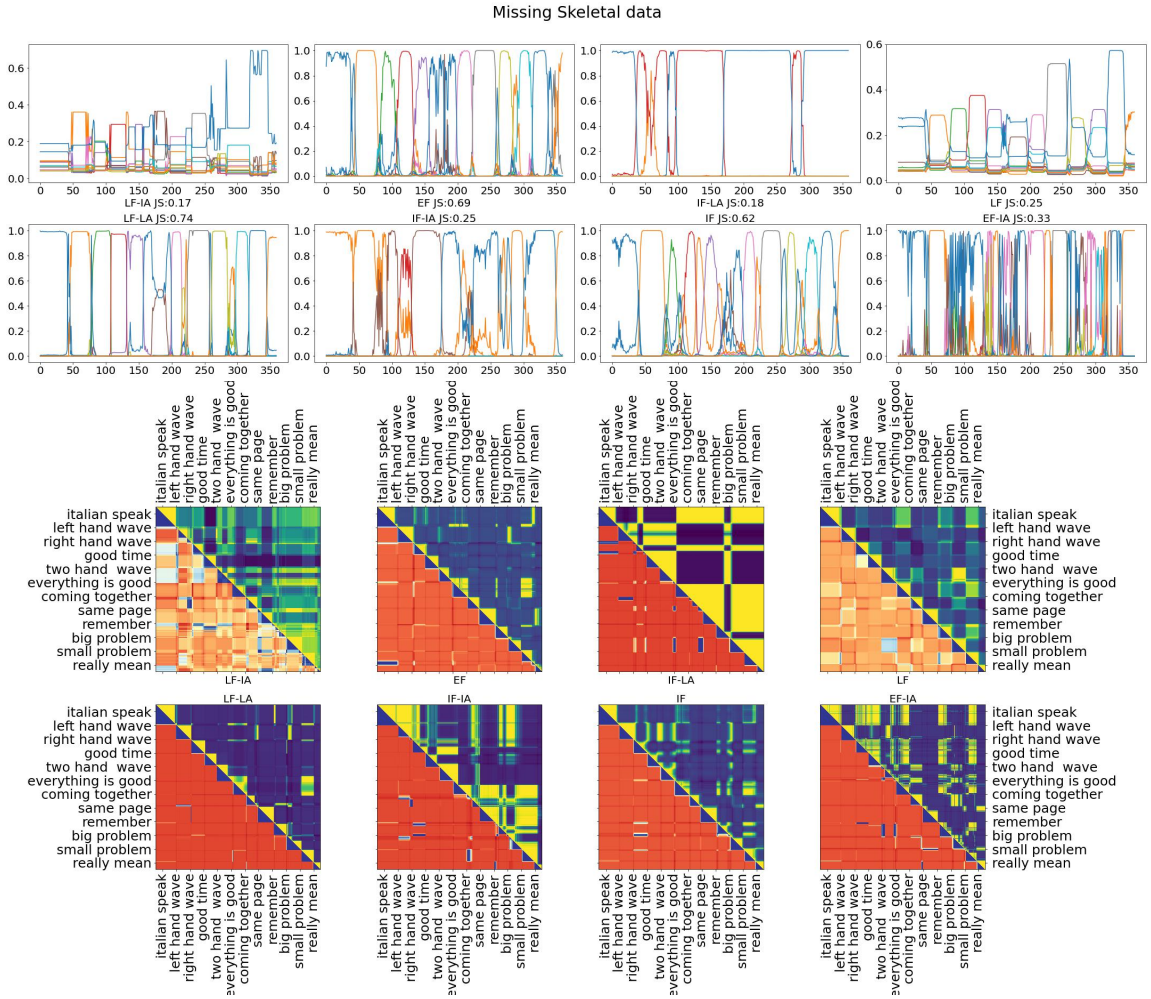

Fig. 8: PPM diagrams consist of a PPM of our reference all modality results (lower triangle) juxtaposed with the relevant PPM under investigation (upper triangle) for skeletal data missing

which resulted in the architecture having the lowest score of all tested architectures.

TABLE VIII: Jaccard index with skeletal mask

G	LF-IA	EF	IF-LA	LF	LF-LA	IF-IA	IF	EF-IA
1	0.0	0.96	0.0	0.276	0.991	0.0	0.911	0.017
2	0.154	0.756	0.083	0.219	0.844	0.493	0.777	0.431
3	0.191	0.677	0.08	0.246	0.817	0.487	0.624	0.498
4	0.195	0.779	0.5	0.323	0.887	0.191	0.538	0.017
5	0.157	0.79	0.5	0.322	0.844	0.235	0.51	0.012
6	0.148	0.208	0.0	0.113	0.253	0.344	0.328	0.115
7	0.242	0.367	0.0	0.093	0.463	0.0	0.365	0.47
8	0.197	0.89	0.0	0.361	0.674	0.0	0.703	0.613
9	0.161	0.719	0.0	0.312	0.786	0.0	0.634	0.5
10	0.0	0.672	0.0	0.118	0.546	0.0	0.494	0.0
11	0.269	0.908	0.5	0.222	0.89	0.493	0.832	0.425
12	0.326	0.507	0.5	0.401	0.941	0.807	0.758	0.817

From Table VIII, we can see that the architecture with the highest Jaccard index when masking the skeletal modality is LF-LA with values ranging from 0.253 to 0.991. On the other hand, the architecture with the lowest Jaccard indices is IF-LA with most values close to or at 0.0.

According to Table IX, the highest Jaccard index is achieved by the architecture LF-LA with a value of 0.745. The architecture EF has a Jaccard index of 0.686, which is the second-highest value in the table. The lowest Jaccard index value is 0.17 for the architecture LF-IA.

TABLE IX: Overall Jaccard index skeletal mask

Jaccard index	Architecture	**Jaccard index**	Architecture
0.17	LF-IA	0.745	LF-LA
0.686	EF	0.254	IF-IA
0.18	IF-LA	0.623	IF
0.251	LF	0.326	EF-IA

Remark 11: *In the circumstance where the* ***sparse spatiotemporal modality*** *was removed, intermediate fusion highly degrades.*

With the removal of the sparse spatiotemporal data Fig. 9 (for instance the radar modality which provides depth information), intermediate fusion experienced difficulties in its assessment and in our case confused most gestures as compared to the all-modality baseline. Late fusion with late attention detection confidence was degraded but maintained a high Jaccard index of 0.802 and was still able to distinguish all 12 gestures as shown in the PPM for LF-LA. LF-IA and EF produced the lowest Jaccard index score with 0.151 and 0.178 respectively. We note that the two architectures performed worst in all test with modality masked. The best score was achieved by LF-LA with 0.802.

From Table X, LF-LA architecture has a Jaccard index of

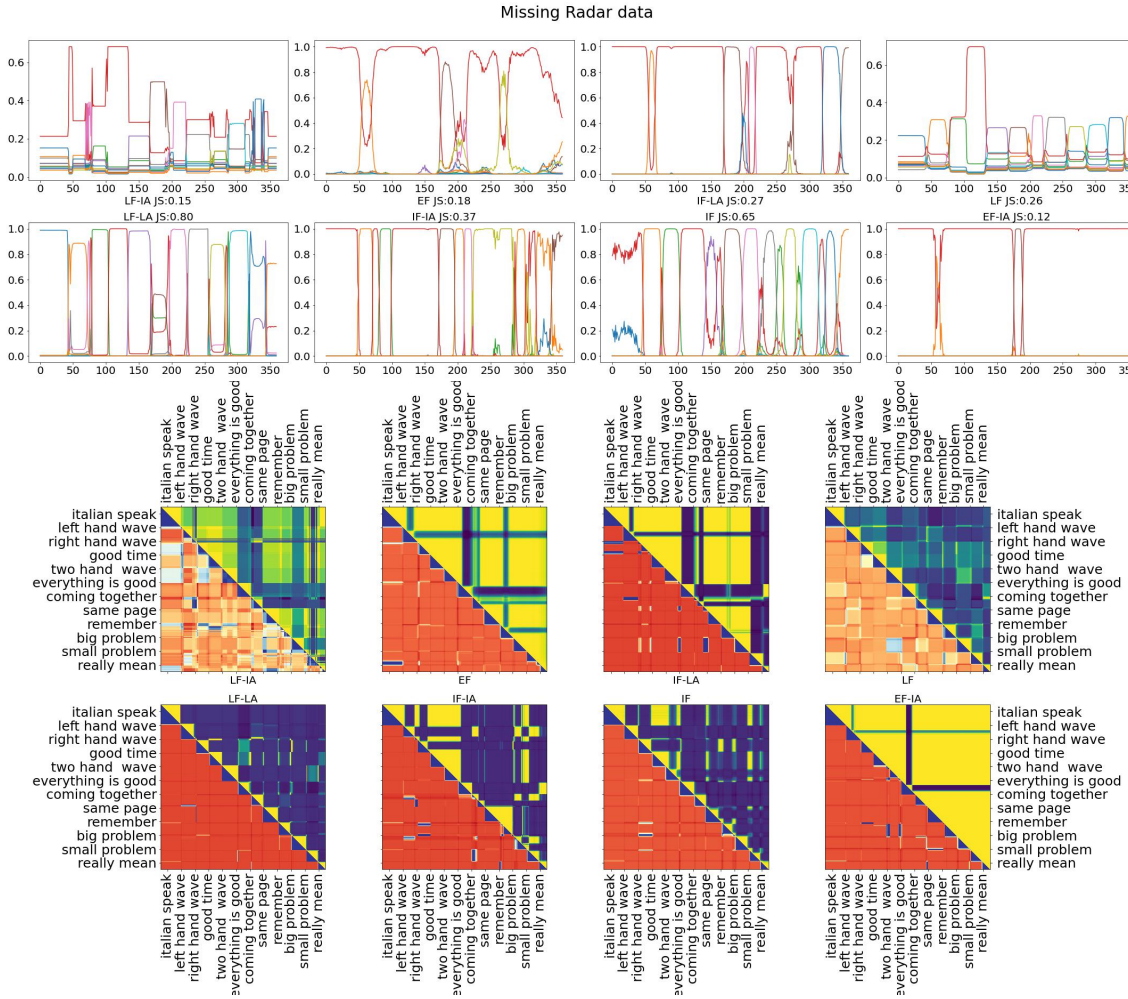

Fig. 9: PPM diagrams consist of a PPM of our reference all modality results (lower triangle) juxtaposed with the relevant PPM under investigation (upper triangle) for missing radar data

TABLE X: Jaccard index with radar mask

G	LF-IA	EF	IF-LA	LF	LF-LA	IF-IA	IF	EF-IA
1	0.0	0.0	0.0	0.224	0.988	0.0	0.0	0.0
2	0.0	0.108	0.116	0.229	0.85	0.33	0.418	0.05
3	0.0	0.197	0.161	0.157	0.708	0.645	0.895	0.037
4	0.341	0.5	0.5	0.348	0.961	0.8	0.914	0.5
5	0.34	0.5	0.5	0.44	0.949	0.5	0.685	0.5
6	0.213	0.139	0.322	0.234	0.62	0.261	0.579	0.116
7	0.349	0.408	0.55	0.209	0.611	0.496	0.853	0.303
8	0.157	0.054	0.25	0.27	0.932	0.399	0.723	0.0
9	0.0	0.116	0.0	0.262	0.831	0.477	0.578	0.0
10	0.119	0.117	0.0	0.236	0.719	0.489	0.705	0.0
11	0.165	0.0	0.32	0.253	0.779	0.0	0.613	0.0
12	0.128	0.0	0.493	0.296	0.679	0.0	0.78	0.0

0.988, which indicates that there is a high level of overlap between the two sets. The EF-IA architecture has the lowest Jaccard indices with the radar mask, with a value close to or at 0.0, indicating the least amount of overlap between the two sets.

The results from Table XI suggest that the architecture with the highest overall Jaccard index for the skeletal mask is LF-LA (0.802) while the architecture with the lowest overall Jaccard index for the radar mask is EF-IA (0.125). Table XIII shows a summary of our findings.

TABLE XI: Overall Jaccard index radar mask

Jaccard index	Architecture	**Jaccard index**	Architecture
0.151	LF-IA	0.802	LF-LA
0.178	EF	0.366	IF-IA
0.268	IF-LA	0.645	IF
0.263	LF	0.125	EF-IA

Remark 12: *There are configurations when the radar modality produces complementary information and others where it produces noise.*

The radar modality can produce complementary information (e.g., depth) to improve the Jaccad index such as in LF-IA, EF and EF-IA and can also produce noise such as in IF-LA, LF, LF-LA and IF as shown in Table XII (OS→OR).

Remark 13: *In our experiments, the merging of radar with optical and skeletal data, following a late merging scheme with late attention produced the best results.*

Remark 13 is supported by Tables VII, IX, XI (summarised in Table XIII) showing better performance for LF-LA with the lowest Jaccard index drop from all modalities with 14.21%, 22.95% and 32.37% for radar, skeletal and optical respectively, showing that the architecture is the most resistant to modality removal. LF-LA in this work consists of pre-trained unimodal models whose

TABLE XII: Difference in Jaccard Index with OSR representing optical, skeletal and radar data respectively

Arch	RS→OR	OS→OR	SR→OS	OSR→OS	OSR→OR	OSR→SR
LF-IA	17.06↑	12.58↓	6.62↑	151.66↓	123.53↓	169.50↓
EF	6.41↑	285.39↓	260.67↓	403.93↓	30.76↓	39.72↓
IF-LA	4.44↓	32.84↑	29.85↑	217.16↓	372.22↓	352.13↓
LF	3.59↑	4.56↑	7.98↑	102.66↓	112.35↓	120.25↓
LF-LA	7.11↑	7.11↑	13.72↑	**14.21**↓	**22.95**↓	**32.37**↓
IF-IA	7.48↓	30.60↑	25.41↑	131.69↓	233.86↓	210.62↓
IF	6.26↓	3.41↑	2.64↓	40.16↓	45.10↓	36.56↓
EF-IA	61.35↑	160.80↓	0.80↓	594.40↓	166.26↓	588.89↓

TABLE XIII: Summary of findings

Opt.	Skel.	Radar	Fusion point	Attention point
✓	✓	✓	EF/IF/LF	LA/IA
	✓	✓	LF	LA
✓		✓	LF	LA
✓	✓		LF	LA

outputs are concatenated and on which attention is applied. The resulting model is trained end-to-end. Attention in this particular configuration is applied to decisions and effectively is not looking at features but relationships between models. Attention in the configuration is used for decision switching and can be used to dynamically switch between different modalities based on their reliability. For example, the model may attend more to the radar modality when the visual input is noisy, and switch back to the visual modality when the radar input becomes noisy. In this way, attention allows the model to dynamically adjust its behaviour, leading to more robust decision-making. Our results align with [38] who found that the late fusion approach with attention mechanism outperforms other baselines in terms of accuracy on a multimodal emotion recognition task. Our results strongly suggest that the late fusion approach with attention mechanism can effectively leverage the complementary information from multiple modalities to improve the accuracy of gesture recognition, making it a promising approach for multimodal fusion in this domain.

Remark 14: *Late attention-based fusion can help recover "hidden" information by selectively attending to the most relevant parts of the fused decision space, which may not be captured by the individual processing of each modality.*

Remark 14 is supported by Fig. 6, 7, 8 and 9 with EF, LF and LF-LA which shows that when early fusion was used, the optical, skeletal and radar modalities were combined at an early stage, which enabled the system to capture the relevant information from the modalities. This resulted in a high Jaccard score. However, when late fusion was used, discriminating information was obscured during the processing stage. We know that this information must be present as it reemerges when attention is applied to the fused decision softmax layer is the final stage of the fusion process.

VI. Conclusion and Future Work

Category information, especially the conflict in category predictions, is difficult to handle in sensor fusion. In this work, we present a novel attention mechanism specifically designed to tackle sparse data. We show that this attention performs better on spatially and temporally sparse data. We also show that where we choose to merge and where we choose to put attention in multimodal fusion with radar can significantly affect the resulting model and subsequent real-time detection. First, we found that multimodal fusion with the attention mechanism of different inputs results in a clear improvement over unimodal approaches due to the complementary nature of the radar modality. From the eight different architectures tested, it was noted that late fusion with late attention suffered the least degradation and outperformed early and intermediate fusion in circumstances where one of the modalities is masked. We also found that late fusion with late attention can also recover "hidden" information from individual softmax from different modalities. We think there is much to be found on the effect of fusion and attention positioning and that further research on the topic is warranted.Our research findings have opened up numerous potential avenues for future exploration and improvement. This work has merely scratched the surface of what is possible in this field. We have identified a few key areas to focus on in our future research efforts. Firstly, we are interested in enhancing our fusion model to manage a wider array of data sources. Our current work has focused predominantly on a limited set of data types. We aim to diversify this in our future research, aiming to include a variety of different sources such as text, images, audio, time-series data, and more complex forms of data such as graph data. By doing so, we hope to construct a fusion model that is more flexible and robust in diverse data scenarios. Secondly, we wish to delve into the potential of emerging deep learning techniques for multimodal fusion. As the field of artificial intelligence progresses, new methodologies, algorithms, and techniques are being developed at an impressive rate. We plan to explore novel techniques such as transformer-based models, self-supervised learning, and neural architecture search, among others. These novel approaches may provide opportunities to improve the efficiency, accuracy, and applicability of multimodal fusion models. Lastly, we aspire to apply our fusion method in practical, real-world settings. In particular, we see a massive potential for our fusion model in the domains of autonomous driving and surveillance. These areas demand robust, reliable models that can handle vast amounts of diverse data in real-time. Applying and testing our model in these high-stake environments will provide valuable insights into its efficacy and practicality.

References

[1] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1685–1699, 2009.

[2] A. Akl, C. Feng, and S. Valaee, "A novel accelerometer-based gesture recognition system," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6197–6205, 2011.

[3] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, and J. Yang, "A framework for hand gesture recognition based on accelerometer and emg sensors," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 41, no. 6, pp. 1064–1076, 2011.

[4] G. Chen, Z. Xu, Z. Li, H. Tang, S. Qu, K. Ren, and A. Knoll, "A novel illumination-robust hand gesture recognition system with event-based neuromorphic vision sensor," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 2, pp. 508–520, 2021.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[6] I. Rodríguez-Moreno, J. M. Martínez-Otzeta, I. Goienetxea, I. Rodriguez, and B. Sierra, "A new approach for video action recognition: Csp-based filtering for video to image transformation," *IEEE Access*, vol. 9, pp. 139 946–139 957, 2021.

[7] Z. Wang, Z. Yu, X. Lou, B. Guo, and L. Chen, "Gesture-radar: A dual doppler radar based system for robust recognition and quantitative profiling of human gestures," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 1, pp. 32–43, 2021.

[8] Y. Gu, X. Zhang, Y. Wang, M. Wang, H. Yan, Y. Ji, Z. Liu, J. Li, and M. Dong, "Wigrunt: Wifi-enabled gesture recognition using dual-attention network," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 736–746, 2022.

[9] B. van Amsterdam, M. J. Clarkson, and D. Stoyanov, "Gesture recognition in robotic surgery: A review," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 6, pp. 2021–2035, 2021.

[10] H. Trinh, Q. Fan, P. Gabbur, and S. Pankanti, "Hand tracking by binary quadratic programming and its application to retail activity recognition," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1902–1909.

[11] J. Wan, G. Guo, and S. Z. Li, "Explore efficient local features from rgb-d data for one-shot learning gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1626–1639, 2016.

[12] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 804–811.

[13] M. Chmurski, M. Zubert, K. Bierzynski, and A. Santra, "Analysis of edge-optimized deep learning classifiers for radar-based gesture recognition," *IEEE Access*, vol. 9, pp. 74 406–74 421, 2021.

[14] Y. Dong, J. Liu, and W. Yan, "Dynamic hand gesture recognition based on signals from specialized data glove and deep learning algorithms," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2021.

[15] J. P. Sahoo, S. P. Sahoo, S. Ari, and S. K. Patra, "Hand gesture recognition using densely connected deep residual network and channel attention module for mobile robot control," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[16] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, T. S. Alrayes, H. Mathkour, and M. A. Mekhtiche, "Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation," *IEEE Access*, vol. 8, pp. 192 527–192 542, 2020.

[17] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *2016 Int. Joint Conf. on Neural Networks (IJCNN)*, 2016, pp. 381–388.

[18] S.-J. Ryu, J.-S. Suh, S.-H. Baek, S. Hong, and J.-H. Kim, "Feature-based hand gesture recognition using an FMCW radar and its temporal feature analysis," *IEEE Sensors Journal*, vol. 18, no. 18, pp. 7593–7602, sep 2018.

[19] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, "Hand-gesture recognition using two-antenna doppler radar with deep convolutional neural networks," *IEEE Sensors Journal*, vol. 19, no. 8, pp. 3041–3048, apr 2019.

[20] J.-W. Choi, S.-J. Ryu, and J.-H. Kim, "Short-range radar based real-time hand gesture recognition using LSTM encoder," *IEEE Access*, vol. 7, pp. 33 610–33 618, 2019.

[21] S. K. Leem, F. Khan, and S. H. Cho, "Detecting mid-air gestures for digit writing with radio sensors and a CNN," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1066–1081, apr 2020.

[22] H. Liu and Z. Liu, "A multi-modal dynamic hand gesture recognition based on radar-vision fusion," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[23] S. Skaria, A. Al-Hourani, and R. J. Evans, "Deep-learning methods for hand-gesture recognition using ultra-wideband radar," *IEEE Access*, vol. 8, pp. 203 580–203 590, 2020.

[24] L. O. Fhager, S. Heunisch, H. Dahlberg, A. Evertsson, and L.-E. Wernersson, "Pulsed millimeter wave radar for hand gesture sensing and classification," *IEEE Sensors Letters*, vol. 3, no. 12, pp. 1–4, dec 2019.

[25] Y. Zhang, S. Dong, C. Zhu, M. Balle, B. Zhang, and L. Ran, "Hand gesture recognition for smart devices by classifying deterministic doppler signals," *IEEE Transactions on Microwave Theory and Techniques*, vol. 69, no. 1, pp. 365–377, jan 2021.

[26] C. Wang, S. Chen, Y. Yang, F. Hu, F. Liu, and J. Wu, "Literature review on wireless sensing-wi-fi signal-based recognition of human activities," *Tsinghua Science and Technology*, vol. 23, no. 2, pp. 203–222, apr 2018.

[27] S. Hazra and A. Santra, "Radar gesture recognition system in presence of interference using self-attention neural network," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, dec 2019.

[28] Y. Hu, Y. Wong, Q. Dai, M. Kankanhalli, W. Geng, and X. Li, "sEMG-based gesture recognition with embedded virtual hand poses and adversarial learning," *IEEE Access*, vol. 7, pp. 104 108–104 120, 2019.

[29] J. F.-S. Lin, M. Karg, and D. Kulic, "Movement primitive segmentation for human motion modeling: A framework for analysis," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 3, pp. 325–339, jun 2016.

[30] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Computer Vision - ECCV 2014 Workshops*, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Cham: Springer International Publishing, 2015, pp. 474–490.

[31] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1583–1597, 2016.

[32] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Tmmf: Temporal multi-modal fusion for single-stage continuous gesture recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 7689–7701, 2021.

[33] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal gesture recognition using 3-d convolution and convolutional lstm," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.

[34] P. Gavrikov, "visualkeras," https://github.com/paulgavrikov/visualkeras, 2020.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[36] P. Towakel, "12 hand gestures," 2022. [Online]. Available: https://www.kaggle.com/dsv/3935841

[37] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[38] J. Lee, J. Kim, Y. Lee, and K. Lee, "Late fusion with attention mechanism for multimodal emotion recognition," in *Proceedings of the 2019 ACM on Multimedia Conference*, 2019, pp. 2221–2229.

Praveer Towakel received his B.Sc. degree in Physics from the University of Mauritius in 2016, and is currently pursuing a PhD within the discipline of Design Engineering at Middlesex University. His current research interests include Gesture Recogni- tion, Radar Systems and Machine Learning.

David Windridge is Professor of Data Science and Machine Learning at Middlesex University, London and leads the Dept. of Computing's Data Science activities. He received his Ph.D from Bristol University, UK in Astrophysics/Cosmology. His current research interests centre on methodological development/application in the areas of machine learning, quantum computing, and cognitive systems. He has played a leading role on a number of large-scale ML projects in academic/industrial settings, having authored around 200 publications. He holds a visiting position at University of Surrey, UK, and sits on the editorial board of Springer *Quantum Machine Intelligence*.

Huan X. Nguyen (M'06–SM'15) received the B.Sc. degree from Hanoi University of Science & Technology, Vietnam, in 2000, and the Ph.D. degree from University of New South Wales, Australia, in 2007. He is a Professor of Digital Communication Engineering at Middlesex University London (U.K.), where he is also the Director of London Digital Twin Research Centre and Head of the 5G/6G I& IoT Research Group. He leads research activities in digital twin modelling, 5G/6G systems, machine-type communication, digital transformation and machine learning within his university with focus on industry 4.0 and critical applications (disasters, intelligent transportation, health). He has been leading major council/industry funded projects, publishing 130+ peer-reviewed research papers, and serving as chairs for international conferences (ICT'19-21, ICEM2021, IWNPD'17, PIMRC'20, FoNeS-IoT'20, ATC'15). He is a Senior Member of the IEEE and a Senior Fellow of the HEA. Prof. Huan X. Nguyen is currently a visiting professor at International School, Vietnam National University, Hanoi, Vietnam.