2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) | 978-1-7281-8470-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICMLA51294.2020.00225

Radar Trajectory-based Air-Writing Recognition using Temporal Convolutional Network

Muhammad Arsalan*†, Avik Santra*, Vadim Issakov†
*Infineon Technologies AG, Neubiberg, Germany
†Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany
{muhammad.arsalan, avik.santra}@infineon.com, vadim.issakov@ovgu.de

***Abstract*—Air-writing systems offer users a virtual board to write characters or words in free space using fingers or hand movements. Several works have been proposed in literature that aim to use different sensors to enable such a system as an alternative to the keyboard and click form of human-machine interfaces. The advancement of miniature radar sensors and deep learning has enabled precise estimation and tracking of finger or marker movement followed by character recognition to offer an effective air-writing solution. However, deviating from earlier works in literature that make use of a network of radars to effectively track and recognize characters, in this paper, we propose to use only one or two radars to sense the local hand trajectory. We propose to use 1D temporal convolutional network (TCN) for simultaneous feature extraction and temporal modeling to recognize the drawn character from the local target trajectory. A dataset with 3750 character instances has been recorded using a 60-GHz millimeter-wave frequency-modulated continuous wave radar (FMCW) radar. We demonstrate the proposed end to end solution achieves a mean accuracy of 99.11% and 91.33% for two radar and one radar-based solution respectively outperforming other deep architectures.**

***Index Terms*—Human-Machine Interface, Character Recognition, Air-Writing, Temporal Convolutional Network.**

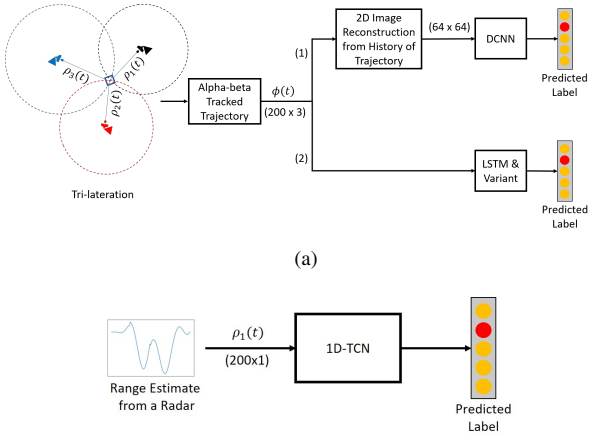


Fig. 1. (a) Conventional Solution using three or more radar in a network, (b) Proposed Solution using one or two radars in the setup. The $\phi(t)$ represents the global trajectory while $\rho_1(t)$, $\rho_2(t)$ and $\rho_3(t)$ represents the local trajectories coordinates of three radars respectively.

I. Introduction

Gesture recognition and character recognition enables an attractive and intuitive form of human-machine interface. Contrary to camera-based solutions, radar offers a promising modality since they are not vulnerable to operating illumination conditions, they can operate through enclosures thus offering an aesthetically appealing solution, can sense minute motions enabling a responsive solution and are privacy-preserving. Radar-based gesture recognition system allow users to make predefined or user-defined motions to control the display or application interface in short-range applications such as smart-phones or smart-watches [1] [2] [3] [4] [5] and far-range applications such as smart TVs or projectors [6][7]. Alternately, character recognition allows users to input alpha-numerical characters drawn in free-space by sensing the trajectory sketched by the hand or marker movement over time. Traditionally, radar-based solutions rely on a network of three or more radars to sense the precise global location (through trilateration [8] [9] or sensor fusion [10]) and track the drawn trajectory to classify the drawn character through deep convolutional neural network or long-short term memory.

In [8], a network of three frequency-modulated continuous wave radar (FMCW) radar in either horizontal or vertical imaginary board configuration is proposed to drawn characters using a hand marker. The target localization and tracking is achieved through least-squares trilateration and $\alpha\beta$ filter, followed by convolutional neural network (CNN) or long short-term memory (LSTM) for recognition of the trajectories into intended characters. In [9] a network of three ultra wide-band (UWB) radars in triangular geometry is proposed to sense and recognize characters drawn in mid-air. In [10], authors use a network of four UWB radars for the air-writing system to localize finger movements in space using extended Kalman filter-based sensor fusion, followed by geometric transformation for robust features extraction for subsequent deep convolutional neural network (DCNN) based classification. However, since all these proposed solutions rely on network of radars, they pose difficulty for deployment in space constraint systems such as monitors and TV screens. Furthermore, the success of trilateration or sensor-fusion based processing is heavily dependent on the intersecting field-of-view (FoV) of the operating radars, which can be limiting due to hand occlusion or physical placements. In this paper, thus we propose a 1D temporal convolutional network (TCN) model to sense the hand marker trajectory using only one or two radars. TCNs are state-less models used for temporal modeling of input data using dilated convolutional neural networks, originally proposed by [11]. TCNs utilize causal dilated convolutions requiring significantly

978-1-7281-8470-8/20/$31.00 ©2020 IEEE
DOI 10.1109/ICMLA51294.2020.00225

less memory compared to LSTMs, which are typically difficult to realize on embedded solutions. Recently, TCNs have been shown to out-perform LSTM and hidden Markov model, in terms of accuracy and compute time, for action recognition [12], speech modeling [13], biomedical data [14] and gesture recognition [15]. In this paper, we explore TCNs for the task of simultaneous feature extraction (equivalent to that of trilateration) and modeling the local radar trajectories from one or two radars for recognition of the drawn characters, thus offering a low-memory footprint, fast inference time and higher accuracy solution. Figure 1 presents the comparison between the traditional solution utilizing a network of (three or more) radars and the proposed solution using 1D TCN to recognize the drawn characters from local trajectories.

The contributions of the paper are as follow -

- An air-writing system using only one or two radars is proposed, thus instead of global location, the local location of the hand or marker is used as input to the classifier.
- A novel 1D TCN architecture is proposed for joint feature extraction and recognition of drawn local radar trajectory.
- The proposed air-writing recognition through local trajectory is demonstrated through *Infineon*'s 60 GHz FMCW radar sensor data paving the way for practical deployment of such solutions in space or cost constraint systems.

II. Background

A. FMCW Radar Principle

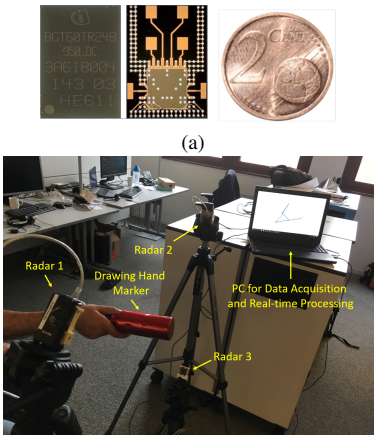
FMCW radar transmits a linearly increasing frequency waveform, called chirp, which after reflection by an object is collected by a receiver antenna. Afterward, both the transmitted and received signal are mixed together by a mixer at the receiver and the resultant low-pass filtered signal is called Intermediate Frequency signal or IF signal. The frequency of the FMCW waveform with bandwidth B and duration T can be expressed as:

$$f_T(t) = f_c + \frac{B}{T}t, \tag{1}$$

where f_c is the ramp start frequency. The reflected signal from the target is mixed with a replica of the transmitted signal resulting in a beat signal. Assuming $\tau_k/T << 1$, the down-converted IF signal due to superposition of K scatterers can therefore be expressed as:

$$s_{IF}(t) = \sum_{k=1}^{K} \exp\left(2\pi\left(\frac{2f_cR_k}{c} + \left(\frac{2f_cv_k}{c} + \frac{2BR_k}{cT}\right)t\right)\right), \tag{2}$$

where $\tau_k = \frac{2R_k + v_k t}{c}$ is the round trip propagation delay between the transmitted and received signal after reflection from the k^{th} target with range R_k and radial velocity v_k. The constant c represents the speed of light. The received signal $s_{IF}(t)$ is first sampled, then the range information of the target is estimated by spectral analysis using fast Fourier transform (FFT) along each chirp, referred to as fast-time FFT. The fast-time complex FFTs along chirps in a frame are combined to boost the signal to noise ratio and referred as coherent



(a)

(b)

Fig. 2. a) *Infineon's BGT60TR24B* FMCW radar chipset. b) Setup of trilateration-based radar air-writing system using network of three radars.

pulse integration. Following the pulse integration, a moving target indicator (MTI) filter is applied to remove the reflections from TX-to-RX leakage or stationary objects in the FoV. In most cases, this MTI filter is implemented through a first-order moving average filter. Following the MTI filter, an energy-based adaptive threshold detector is used to detect the start and stop of the drawn character. Once detected, the local range is estimated as the peak from the pulse integrated and MTI filtered fast-time FFT spectra.

B. Trilateration based Air-Writing

Once the range of the target, in this application the hand or marker, w.r.t. each of the radar sensors in the network is estimated the next step is to feed the range estimates into the trilateration algorithm. Trilateration is a technique that estimates the global coordinates of the target by utilizing the local range estimates to three or more reference points/radars, whose coordinates are known apriori. To illustrate the trilateration technique, consider three reference points, namely $R_1(x_1, y_1, z_1)$, $R_2(x_2, y_2, z_2)$ and $R_3(x_3, y_3, z_3)$. Let ρ_1, ρ_2 and ρ_3 denote the distance of the target $T(px, py, pz)$ from reference points R_1, R_2 and R_3 respectively. Then the coordinates of the target point are given by the following set of equations:

$$\begin{cases} (px - x_1)^2 + (py - y_1)^2 + (pz - z_1)^2 = \rho_1^2 \\ (px - x_2)^2 + (py - y_2)^2 + (pz - z_2)^2 = \rho_2^2 \\ (px - x_3)^2 + (py - y_3)^2 + (pz - z_3)^2 = \rho_3^2. \end{cases} \tag{3}$$

After algebraic modifications, this can be expressed in matrix-vector form as:

$$\begin{bmatrix} 1 & -2x_1 & -2y_1 & -2z_1 \\ 1 & -2x_2 & -2y_2 & -2z_2 \\ 1 & -2x_3 & -2y_3 & -2z_3 \end{bmatrix} \begin{bmatrix} px^2 + py^2 + pz^2 \\ px \\ py \\ pz \end{bmatrix} = \begin{bmatrix} \rho_1^2 - x_1^2 - y_1^2 - z_1^2 \\ \rho_2^2 - x_2^2 - y_2^2 - z_2^2 \\ \rho_3^2 - x_3^2 - y_3^2 - z_3^2 \end{bmatrix}. \tag{4}$$

Above eq. (4) can be simply written in the form:

$$\mathbf{Ax} = \mathbf{b}, \tag{5}$$

Subject to the constraint $x \in C$, where,
$C = \{(x_0, x_1, x_2, x_3)^T \in \mathbb{R}^2 / x_0 = x_1^2 + x_2^2 + x_3^2\}$.

The above problem is tractable and its solution has been provided in [16] and has been utilized in this paper for generating the reference trajectories used as baseline [8] to compare our proposed solution.

After the global coordinates are estimated, $\alpha\beta$ filter is applied to smoothen the trajectory data over space to avoid perturbations due to noise. The $\alpha\beta$ filter iterates through predictions and updates to smoothen the trajectory. The prediction step process is expressed as follows:

$$\begin{aligned} \phi(n) &= \bar{\phi}(n-1) + Tv(n-1), \\ v(n) &= \bar{v}(n-1), \end{aligned} \tag{6}$$

where T is the measurement update interval or frame time, $\phi(n)$ and $\bar{\phi}(n)$ are the predicted and smoothened global target position at time $t = nT$, respectively, and $v(n)$ and $\bar{v}(n)$ are the predicted and smoothened target velocities at time $t_n = nT$, respectively.

The update process is defined as:

$$\begin{aligned} \bar{\phi}(n) &= \phi p(n) + \alpha(\hat{\phi}(n) - \phi(n)), \\ \bar{v}(n) &= v(n) + \frac{\beta}{T}(\hat{\phi}(n) - \phi(n)), \end{aligned} \tag{7}$$

where $\hat{\phi}(n)$ is the measured position of the target at nT trajectory instant, and α, β are fixed weighting factors.

Once the smoothened global trajectory of the detected character is generated, they are fed as time-series trajectories into LSTM models or a 2D image can be reconstructed, which is then fed into the DCNN model for character classification as depicted in fig. 1(a).

C. Temporal Convolutional Networks

TCN inspired by recent convolutional architectures for sequential data are a modeling approach for time series proposed by Lea [11]. TCN uses dilated convolutional neural networks and has been used in many tasks including speech modeling [17] and human action recognition [18]. The TCN has been designed from two principles; dilated convolution and casual convolution. The casual convolution means that there is no information leakage from future to past i.e. for the prediction of any time step no inputs from the future are considered. This is needed for real time prediction scenarios where only the input from the past and present are available. The size of the receptive field in casual convolutions increases linearly with every additional layer as it can only look at history with size linear to the depth of the network. To overcome this problem, TCN uses dilated convolutions i.e. to stretch out the convolutional kernels such that the receptive field of the kernel increases. The dilation convolution, given an input sequence $x \in \mathbb{R}^T$ and a filter $h : \{0, \cdots, k-1\} \rightarrow \mathbb{R}$ is given as:

$$H(x) = (x *_d h)(x) = \sum_{i=0}^{k-1} f(i) x_{s-d.i}, \tag{8}$$

where $d = 2^\ell$ is the dilation factor, with ℓ the level of the network and the term $x_{s-d.i}$ represents the direction of the past. Using larger dilation enables an output at the top level to represent a wider range of inputs, thus effectively expanding the receptive field of a CNN. Figure 3 shows the data flow of the TCN as used in this work.

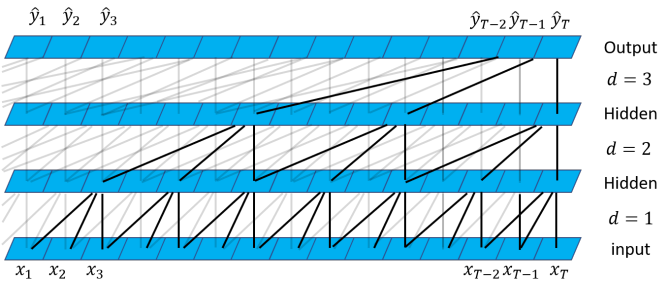


Fig. 3. Data flow in TCN.

III. Proposed Solution

In the proposed approach, we do not utilize the information from all reference radar required to estimate the global coordinates. We use strictly less than three radars and thus eq. (4) is an undetermined system with infinite or no solution. However, through deep learning models, we utilize the temporal range change information to approximate the global data required for the correct classification of the drawn characters. Thus, we explore the full capabilities of the neural network, which can explicitly learn the features directly from the range and range change data in an undetermined setting.

The local coordinate trajectory $\rho_1(t)$ represents the range with time t for a drawn character. Correspondingly, $\phi(t)$ represents the global coordinate trajectory $\big(\phi_x(t), \phi_y(t), \phi_z(t)\big)$. Let $f(.)$ be the function that maps or transform the local trajectory $\rho_1(t)$ to global trajectory $\phi(t)$ and $f(.)$ is a smooth function. Using second-order Taylor series expansion about $t = 0$, the global trajectory can be approximated as:

$$\begin{aligned} \phi(t) = f\big(\rho_1(t)\big) = f\big(\rho_1(0)\big) + f'\big(\rho_1(0)\big)\big(\rho_1(t) - \rho_1(0)\big) \\ + \frac{f''\big(\rho_1(0)\big)}{2}\big(\rho_1(t) - \rho_1(0)\big)^2, \end{aligned} \tag{9}$$

where

$$f'(\rho_1(0)) = [\frac{\partial \phi_x(t)}{\partial \rho_1(t)}\partial\rho_1(t), \frac{\partial \phi_y(t)}{\partial \rho_1(t)}, \frac{\partial \phi_z(t)}{\partial \rho_1(t)}]_{t=0},$$

and

$$f''(\rho_1(0)) = [\frac{\partial^2 \phi_x(t)}{\partial r^2(t)}, \frac{\partial^2 \phi_y(t)}{\partial r^2(t)}, \frac{\partial^2 \phi_z(t)}{\partial r^2(t)}]_{t=0},$$

and $t \in \tau_i$ where τ_i is the i^{th} time segment over which the Taylor expansion is valid. Note that if τ_i is degenerate, the

function $f(.)$ does not have unique solution. The multi-class classification problem can then be expressed as:

$$c_p = g_\theta\Big(\{\phi(t \in \tau_1), \phi(t \in \tau_2), \cdots, \phi(t \in T)\}\Big), \qquad (10)$$

where c_p denotes the prediction of the drawn character, g_θ is the neural network function with parameters θ and T is the cardinality of the time segments to piece-wise approximate the entire trajectory into fixed but unknown $\{f_i'(\rho_1(0)), f_i''(\rho_1(0))\}_{i=i}^{i=T}$. In this paper, we propose to achieve this operation $g_\theta(.)$ using TCN and compare our results with LSTM and 1D CNN-LSTM based solutions.

Figure 1 presents the comparison between the conventional solution and the proposed solution. The trilateration, $\alpha\beta$ tracking filter and track history functionalities in the conventional pipeline are replaced by 1D TCN for temporal modeling of the local trajectories. The TCN extracts the features through temporal modeling per character, from which the classification is done through fully-connected softmax. The classification in the conventional pipeline is achieved through 2D-DCNN or LSTM.

A. Architecture

The architecture takes an input of 200×1 where 200 is the time steps and 1 represents the feature (2 in case of two radar sensors) i.e. each sensor contributes to one channel. These features are fed to TCN which is composed of dilation factors increasing exponentially to combines features from different time steps and stored them in a single feature vector. In the proposed setup, 1 time step is consider at one time which means that only one range value (two in case of two radars) or sensor reading goes to the TCN at one time step. The TCN is composed of 9 layers of dilated causal convolution with dilation factors respectively as $[2^0, 2^1, \ldots, 2^8]$. For each layer, 20 filters have been used and the kernel size of 6 and with stride set to 1. Each dilated causal convolution layer is followed by weight normalization, ReLU activation and dropout of 0.3 rate have been appended during training. The output layer of the network is a fully connected layer appended with softmax function.

The weights w for 2D convolutional layers were drawn randomly from a uniform distribution between $[-limit, limit]$, where $limit = \sqrt{\frac{6}{N_{in}+N_{out}}}$ and N_{in}, N_{out} are the number of input and output units, respectively. Whereas the weights for the Dense layer were initialized by drawing randomly from a normal distribution $N(0, 0.01)$ and for biases normal distribution with mean 0.5 has been used. For optimizer we have used adaptive moment estimation (Adam) optimizer where the the learning rate (alpha) is set to 0.002 and the exponential decay rate for the first (beta1) and second (beta2) moment estimate are set to 0.9 and 0.999. For numerical stability the epsilon is set to $1e-8$.

B. Loss Function

We have used the most common loss function used for multi-class classifications task known as cross-entropy which is mathematically expressed as:

$$L(\hat{Y}_i, Y_i) = -\sum_{j=1}^{c} y_{ik} log(p_{ij}), \qquad (11)$$

where Y_i is one-hot encoded target vector $(y_{i1}, \cdots, y_{ic})$ and where $\hat{Y}_i = (\hat{y}_{i1}, \cdots, \hat{y}_{ic})$ is the classifier output. The term y_{ij} is one if the i^{th} element is in class j otherwise it is 0. The term $p_{i,j}$ represents indicating whether class label k is the correct classification or not.

IV. Results and Discussion

A. Radar Chipset

For the realization of the proposed system we have used Infineon's *BGT60TR24B* FMCW radar chipset as shown in fig. 2(a). The functional block diagram of the chipset is shown in fig. 2(b). The chipset operates in frequencies ranging from 57 to $64\,\mathrm{GHz}$ with four fully differential receivers and two fully differential transmitters. The chip is bunked in an embedded wafer level ball grid array package with six integrated patch antennas realized with a metal redistribution layer. The voltage-controlled oscillator (VCO) generates highly linear frequency chirps which is facilitated by tune voltage varying from 1 to $4.5\,\mathrm{V}$. The chipset is contained with a programmable frequency divider with two-division ratios to enable the use of both hardware and software phase-locked loop (PLL) systems. The chipset has two transmit antennas and four receiving antennas. Each transmitter channel is realized using a differential cascade stage and has a gain of about $6\,\mathrm{dBi}$. The receiving antennas have a combined gain of $10\,\mathrm{dBi}$.

The chipset *BGT60TR24B* is configured with the system parameters and derived parameters provided in tab. I. The same system parameters are used for the air-writing system using a sparse network of radars proposed in this paper.

TABLE I
Operating Parameters.

Parameters with symbol	Value
Range resolution of single radar	$2.5\,\mathrm{cm}$
Elevation θ_{elev} per radar	$70°$
Azimuth θ_{azim} per radar	$70°$
Ramp start frequency, f_{min}	$57\,\mathrm{GHz}$
Ramp stop frequency , f_{max}	$63\,\mathrm{GHz}$
Bandwidth, B	$6\,\mathrm{GHz}$
Chirp time, T_c	$171.2\,\mu\mathrm{sec}$
Sampling frequency, fs	$0.747\,664\,\mathrm{MHz}$
Number of samples per chirp, Ns	128

B. Dataset and Setup

For training and evaluation of our proposed setup, an alphanumeric air-written dataset has been created containing capital alphabets from A-J and numerals 1-5. This dataset has been obtained in two scenarios: one with a vertical virtual board and another with a horizontal virtual board as shown in fig. 2(b). As soon as the marker enters the radar FoV and is detected by all the three radars represents the start point of the trajectory, the frame counting starts. The end of the trajectory

or character is achieved by dropping the marker out of the FoV when the character is written in the virtual board.

It is to be noted the experimental setup shown is of the conventional approach based on trilateration. Each senors ADC data is fed to PC via a USB device. The data has been automated and starts with the hand mark detection and ends when the hand marker is dropped down from the radar FoV and is not detected for 20 consecutive frames. Each character and numeral in the dataset is recorded and labeled with large inter-class variance. For each character, we select 200 samples for training and 50 samples for testing using random trials.

C. Classification

To evaluate the classification performance of the proposed system average accuracy has been used as a performance measure. The average accuracy of both single radar and two radars is shown in tab. II. It can be seen that the proposed method with two radars achieves a similar level of accuracy as is achieved by state-of-the-art trilateration-based approach [8].

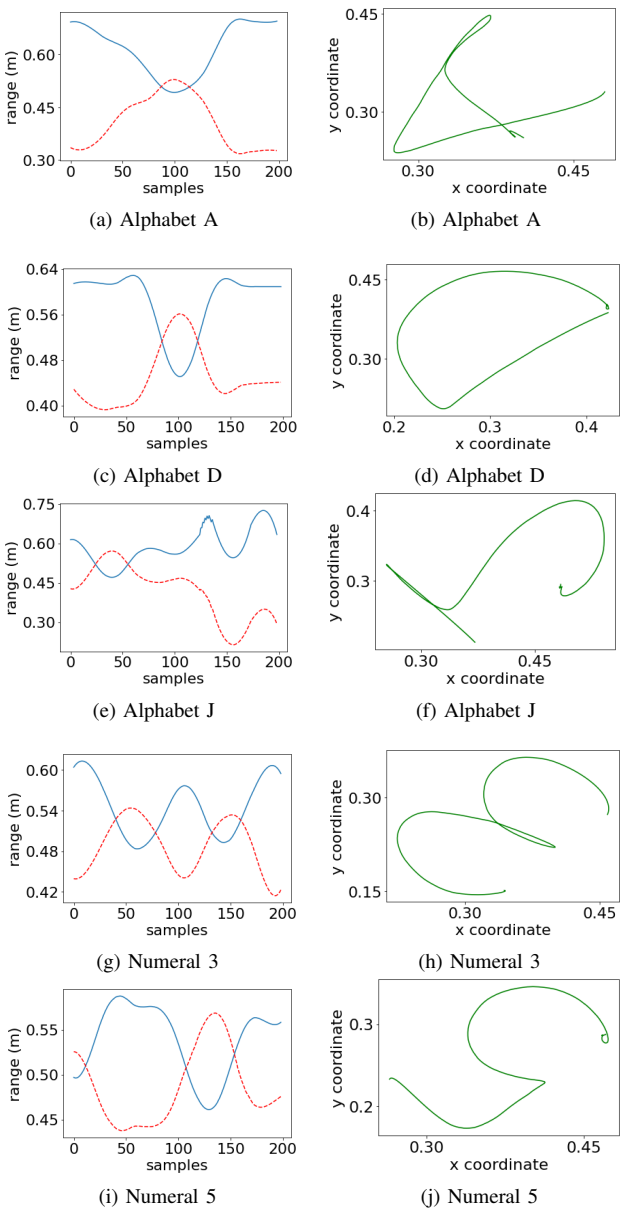
TABLE II
CLASSIFICATION ACCURACY OF THE PROPOSED SOLUTION TO OTHER MODELS AND BASELINE USING THREE RADARS.

Approach	No. of radars	Trajectory	Accuracy (%)	Memory/Parameters
ConvLSTM [8]	3	Global	98.33	1.56 MB / 133,903
DCNN [8]	3	Global	98.33	752 kB / 173,631
LSTM	1	Local	89.32 ± 6.66	1.10 MB / 93,465
LSTM	2	Local	93.33 ± 5.32	1.10 MB / 94,065
1D CNN-LSTM	1	Local	90.33 ± 4.44	891 kB / 94,139
1D CNN-LSTM	2	Local	97.33 ± 2.67	891 kB / 94,139
Proposed 1D TCN	**1**	**Local**	$\mathbf{91.33 \pm 4.66}$	**643 kB/ 41,635**
Proposed 1D TCN	**2**	**Local**	$\mathbf{99.11 \pm 0.89}$	**644 kB / 41,775**

D. Discussion

An air-writing system using only one or two radar systems with TCN model is proposed that is capable of simultaneous learning implicit features and temporal modeling directly from the local radar trajectory. This particularly is advantageous in space-constraint devices such as personal monitors or screens and offers a more cost-effective solution. This is achieved by simultaneous feature extraction and temporal modeling of the trajectory's in correlation to its earlier measurements using TCN. The temporal modeling helps in finding the solution to otherwise under-determined problem. Furthermore, in [8] only single character recognition is demonstrated, here the proposed system is extended to detect words and continuous numerals. To further illustrate the performance of the proposed method we have compared it with other deep temporal models such as LSTM and 1D CNN-LSTM and demonstrated that our proposed system out-performs them in terms of classification accuracy. Further, the proposed architecture has a small memory footprint (approx. 644 kB without quantization, pruning and fusion) and requires less training parameter (41,635) compared to other models, making it an efficient and suitable method for commodity hardware.

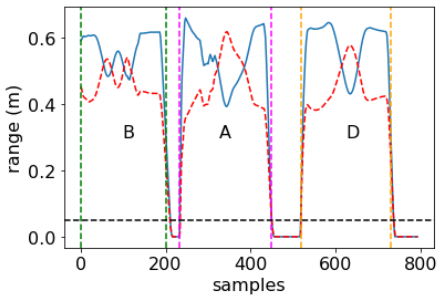
It is evident from tab. II that the proposed system achieves similar level, $99.11 \pm 0.89\%$, of accuracy as is achieved by state-of-the-art method [8] classification 98.33%. This improvement



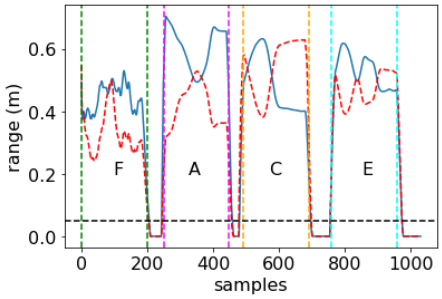
(a) Alphabet A (b) Alphabet A (c) Alphabet D (d) Alphabet D (e) Alphabet J (f) Alphabet J (g) Numeral 3 (h) Numeral 3 (i) Numeral 5 (j) Numeral 5

Fig. 4. 1st column shows the local trajectories of the character. In case of one radar, only blue curve reading is fed to the system, while in case of two radars both the blue and red curves are fed to the system. While the final global trajectories that are obtained with trilateration are shown in column 2 respectively.

is achieved partly due to replacing the hand-crafted feature extraction (i.e. trilateration), which is general is prone to inaccuracies arising due to occlusion of fingers on one or more radar and missed detection of the target on one of the radars leading to unreliable global coordinates. The proposed solution is immune to such inaccuracies as the features along with the temporal modeling are jointly learned intrinsically by the model. It can be seen in the tab. II the accuracy drops down by $4-8\%$ when single radar is used due to much fewer dimensions available for the model to learn the sufficient features which can be overcome by a faster frame-rate of the radar system or using Doppler trajectory information. Figure 4 shows the change of



(a) Word BAD



(b) Word FACE

Fig. 5. a) The local trajectories for the word "BAD" whereas b) shows the local trajectories for the word "FACE". In case of single radar, only blue curve is fed to the system, while in case of two radars both the red and blue curve are fed to the system.

range from each radar for a few alphabets and numerals for both the radars (blue and red line). The global coordinates of the character and numerals are as well shown, which are obtained by feeding the range information from three radars and then using trilateration to reconstruct the global trajectory.

Figure 5 shows the results of continuous character writing. The characters are drawn in a consecutive manner with certain delays. The delay between two consecutive characters is not fixed and varies with the user's interaction. To identify the start and end of the character we used thresholding as outlined earlier while the character is drawn over time. For example, fig. 5 shows two words BAD and FACE written in a continuous fashion with the segmentation of the individual word. The black dotted horizontal line indicates the minimum range threshold set when the target is in the FoV or virtual board the range values obtained with radar are higher enough as compared to when the target is outside. Thus, the simple thresholding technique was quite effective and was evaluated on more than 50 words ranging from 2 to 4 characters. The evaluation accuracy was 100% for the segmentation.

V. Conclusion

Air-writing offers an intuitive and natural mechanism to input characters or numerals in a wide range of human-machine interface applications. Traditional air-writing systems using three or more radars have poor acceptance in practical deployment mainly due to practical limitations associated with placement of the radars in monitor or TV screens and associated costs. Instead of the global trajectory, in this paper, we propose to use the local range trajectory from one or two radars to sense and identify the drawn characters in the virtual board. A novel 1D TCN classifier model has been proposed to effectively model the temporal local trajectory and classify the drawn character. We compare the classification accuracy and memory footprint of our proposed solution to other deep models and trilateration-based approaches. This work is believed to pave the way for practical deployment of such air-writing systems.

References

[1] X. Cai, J. Ma, W. Liu, H. Han, and L. Ma, "Efficient convolutional neural network for fmcw radar based hand gesture recognition," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 2019, pp. 17–20.

[2] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Short-range fmcw monopulse radar for hand-gesture sensing," in *2015 IEEE Radar Conference (RadarCon)*. IEEE, 2015, pp. 1491–1496.

[3] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–19, 2016.

[4] S. Hazra and A. Santra, "Robust gesture recognition using millimetric-wave radar system," *IEEE sensors letters*, vol. 2, no. 4, pp. 1–4, 2018.

[5] A. Santra and S. Hazra, *Deep Learning Applications of Short Range Radars*. Artech House, 2020. [Online]. Available: https://books.google.de/books?id=Qb-VzQEACAAJ

[6] X. Lou, Z. Yu, Z. Wang, K. Zhang, and B. Guo, "Gesture-radar: Enabling natural human-computer interactions with radar-based adaptive and robust arm gesture recognition," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2018, pp. 4291–4297.

[7] 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, 2019, pp. 1409–1414.

[8] M. Arsalan and A. Santra, "Character recognition in air-writing based on network of radars for human-machine interface," *IEEE Sensors Journal*, vol. 19, no. 19, pp. 8855–8864, 2019.

[9] 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, 2020.

[10] F. Khan, S. K. Leem, and S. H. Cho, "In-air continuous writing using uwb impulse radar sensors," *IEEE Access*, vol. 8, pp. 99 302–99 311, 2020.

[11] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[12] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE, 2017, pp. 1623–1631.

[13] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.

[14] D. Jarrett, J. Yoon, and M. van der Schaar, "Dynamic prediction in clinical survival analysis using temporal convolutional networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 424–436, 2019.

[15] M. Scherer, M. Magno, J. Erb, P. Mayer, M. Eggimann, and L. Benini, "Tinyradarnn: Combining spatial and temporal convolutional neural networks for embedded gesture recognition with short range radars," *arXiv preprint arXiv:2006.16281*, 2020.

[16] A. Norrdine, "An algebraic solution to the multilateration problem," in *IEEE IPIN*, 2012.

[17] A. oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," 09 2016.

[18] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," 07 2017, pp. 1623–1631.