RF Micro-Doppler Classification with Multiple Spectrograms from Angular Subspace Projections

Emre Kurtoğlu, *Student Member, IEEE,* Ali C. Gurbuz, *Senior Member, IEEE,* Evie Malaia, Darrin Griffin, Chris Crawford, and Sevgi Z. Gurbuz, *Senior Member, IEEE*

***Abstract*—Radio Frequency (RF) sensors present distinct advantages over cameras or wearables for hand gesture recognition providing high resolution radial range and velocity measurement, being able to operate in dark and through the objects with high temporal and frequency resolutions. Moreover, the flexibility of the complex formatted data allows users to develop their own algorithms to generate various data representations such as time-frequency (Micro-Doppler - μD) maps, or range-Doppler or -angle as a function of time. However, conventional μ-D generation does not regard the angular information of the multiple targets existing in the RF data. Hence, multiple targets with different μ-D signatures at various angular positions create a mixed spectrogram output reducing recognition performance. This paper proposes an angular projection approach on radar data cubes (RDCs) to generate raw radar data for defined angular subspaces. Hence multiple μ-D spectrograms for each angular subspace can be constructed from the projected data. The proposed approach has been tested on RF data for gross body movement and American Sign Language (ASL) recognition. It has been showed that the utilization of angular projected spectrograms increases classification accuracy for ASL and achieves recognition accuracy of 92.6% for 20 word ASL signs.**

***Index Terms*—Micro-Doppler spectrogram, American Sign Language, RF sensors, deep neural networks**

I. Introduction

Recognition of human activities, hand gestures and American Sign Language (ASL) signs have gained an enormous interest in the research community with the development of low cost, low power and small size sensors. RGB [1] and RGB-D cameras [2], motion capture (MOCAP) systems [3], sensor augmented gloves [4] and radars [5], [6] can be listed as most suitable and commonly used sensor types for these sort of applications. Although sensor augmented gloves yield high recognition rates for hand gesture classification tasks, their restrictive nature raises concerns about usability of them in a daily living scenario [7]. Video based applications, on the other hand, offer non-contact sensibility of the environment in a 2-D plane. Addition of extracting depth information capability to the optical cameras provide 3-D localization of the key points on the body or the hands. However, video-based systems inherently invades the privacy of the users as they collect visual imagery of the scene and people around when they are used for indoor sensing applications. Radio Frequency (RF) sensing started to serve as a new, non-invasive, high resolution (in terms of range, velocity, angle, and time) modality which can accurately capture human kinematics [8], and transitional features of moving targets such as hand gestures [9]. RF sensing has also been proven to be very effective in various tasks such as fall detection [10], heart rate [11] and respiration measurement [12], [13], gait abnormality detection [14], [15], and extraction of linguistic features of sign languages [16].

Classical RF based activity recognition utilizes time-frequency representations such as spectrograms. This approach generally assumes existence of a single target/activity class to be recognized. When there are multiple targets present at different ranges or angles, their signatures are generally mixed reducing classification performance. Multiple targets can be separated on range-Angle or range-Doppler maps however these representations don't directly result each targets time-frequency representation. Being able to create spectograms of each target at various angles creates additional capability for recognition of the activity. For example for ALS based applications extraction of linguistic features is a significant task in a sense that they can reveal crucial information about characteristics of the signs described in the ASL-LEX database [17] such as major and minor location, movement type, one/two handedness, and empirical angle of arrival (AoA). While majority of these features can be learned from different RF data representations by training a sophisticated learning model for the task of interest without performing a separate angle estimation step, embedding of the angle information into the input to be used in the learning model can contribute to creation of a more enriched features space, hence resulting in more accurate estimations. In this end, this paper proposes a projection based method to project the raw radar datacube to various angular subspaces resulting specific radar data cubes for each angular subspace. By this way micro-Doppler (μ-D) spectrograms [18] for each angle subspace can be computed, which puts more emphasis (i.e., strengthens) on the μ-D signatures of the targets located within the projected angle interval, and fades out (i.e., weakens) the μ-D signature components of the targets located outside the projected angle interval. In this way, AoA information is being embedded to newly generated μ-D spectrograms besides velocity and time that will enhance classification accuracy for each target or for an activity distributed in angle.

Deep neural networks (DNNs) have been demonstrated to be quite successful in a wide range of tasks from object detection [19] to natural language processing [20]. One important characteristic of DNNs is that they heavily rely on amount and diversity of the presented data to learn the underlying relationships between inputs and outputs. For this reason, designing a customized architecture for a given task is very common and needed. In this work, in order to take

TABLE I: Selected 20 ASL Words

YOU	HELLO	WALK	DRINK	FRIEND
KNIFE	WELL	CAR	ENGINEER	MOUNTAIN
LAWYER	HOSPITAL	HEALTH	EARTHQUAKE	BREATHE
HELP	PUSH	GO	COME	WRITE

advantage of the projection method in the classification stage, two spectrograms are generated from each RF data, for two angular subspace projections, and fed into a multi-branch DNN model.

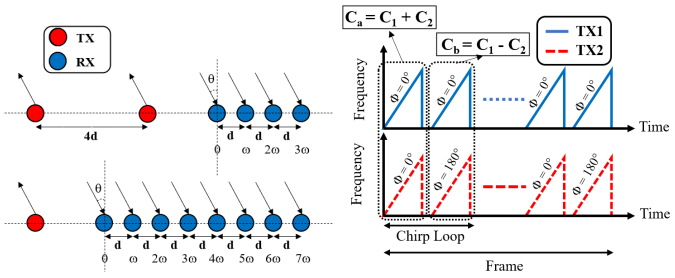
This paper introduces an angular projection approach to embed the angular information captured by the multiple input multiple output (MIMO) radar system into the μ-D spectrograms without requiring computationally intense super-resolution angle estimation methods. In addition a multi-branch DNN model is designed to fuse the feature spaces of left and right projected spectrograms. It is shown that the proposed method can classify 20 ASL signs with over 92% accuracy. In § II, experiment design and testing environment is described along with the acquired dataset. While § III explains the projection method and presents the similarity results of projected spectrograms, § IV describes the DNN model used for training and presents the classification results along with the recognition rates of each sign. Finally § V concludes the paper and discusses the future work.

II. Data Collection and Experimental Setup

In this work, Texas Instrument's AWR1642BOOST 77 GHz frequency modulated continuous wave (FMCW) MIMO automotive radar is coupled with a DCA1000EVM raw data capture card for data acquisition. As an initial study of the proposed projection method, a few samples of human walking and hand moving data is acquired. In the main dataset, 20 high frequency ASL words are selected from the ASL-LEX database while preserving the phonological diversity. A complete listing of the selected words are provided in Table I. 6 non-native (i.e., hearing) participants were hired to conduct the experiment, and in total 945 samples were collected for 20 words (i.e., ∼45 samples per class). Although the participants were not native ASL users, they were trained with native signers' videos before the data collection until they are comfortable to enact all the signs. The participants were seated on a bar stool located 1.5m away from the radar within the direct line of sight. The signs were prompted to the participants through a monitor facing towards them. The dataset is split into 80% and 20% portions for training and testing, respectively.

III. Projection of Radar Data Cubes

The raw data acquired by each channel of the RF sensor are in the format of in-phase (I) and quadrature (Q) samples in a time stream. Having multiple receive and transmit channels allows users to extract angular information (i.e. direction of arrival) in addition to range and velocity.



(a) Building a virtual ULA with 8 receivers.

(b) Configuration of the chirps for BPM.

Fig. 1: Illustration of the working principle and configuration of BPM.

A. Binary Phase Modulation Configuration

The employed 77 GHz FMCW MIMO radar hosts 2 transmitter (TX) and 4 equally spaced receiver (RX) antennas, which forms a uniform linear array (ULA). In order to increase the number of RX channels, hence to have a better angular resolution, one can apply different modulation techniques on the chirp configuration of the transmitted waveform such as time-division multiplexing (TDM) or binary phase modulation (BPM) which enables users to create a virtual array with 1 TX and 2×4=8 RX antennas.

Although TDM-MIMO scheme is easy to implement it does not benefit the complete transmission capabilities of the device, since only one TX is being active at a time. BPM, on the other hand, is centered on modulating the phases of the chirps in a coherent processing interval (CPI), which enables simultaneous transmission from multiple TX antennas while still maintaining separation of these signals. To create the aforementioned virtual array, two different chirp configurations are defined. In the first chirp, C_a, both TX antennas (i.e. TX1 and TX2) transmit the chirp with phase of $\phi°$, while in the second chirp, C_b, TX2 transmits with phase of 180°. The chirps transmitted by both antennas can then be retrieved by $C_1 = (C_a + C_b)/2$ and $C_2 = (C_a - C_b)/2$ for TX1 and TX2, respectively. While Figure 1a depicts the actual and the obtained virtual array, Figure 1b shows the chirp configuration of the BPM scheme for each TX antenna.

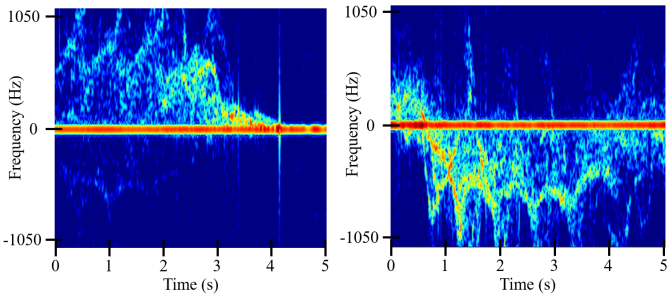
The angular resolution, θ_{res}, of the system can be computed by:

$$\theta_{res} = \frac{\lambda}{M \times d \times cos(\theta)} \tag{1}$$

where λ is the wavelength, M is the number of channels, and d is the distance between consecutive RX antenna elements. Thus, increasing M from 4 (real) to 8 (virtual) yields doubly finer angular resolution.

The phase difference, ω, between RX channels for a target located at $\theta°$ can then be computed as:

$$\omega = \frac{2\pi d sin(\theta)}{\lambda} \Rightarrow \theta = sin^{-1}(\frac{\lambda\omega}{2\pi d}) \tag{2}$$



(a) Walking towards radar. (b) Walking away from radar.

Fig. 2: μ-D spectrograms of RF data.

B. Steering Matrix

A steering vector, a, which represents the set of phase delays for an incoming ray at each virtual RX element, can be constructed as follows:

$$a(\theta) = [1 \; e^{-j(2\pi f_c d \sin(\theta)/c)} \; ... \; e^{-j(2\pi f_c (M-1) d \sin(\theta)/c)}]^T \tag{3}$$

where f_c is the center frequency and c is the speed of light. Repeating a for each θ_i where $\theta_i \in$ [-90° 90°], one can obtain a $M \times N$ steering matrix, $\mathbf{A}$, where N=181.

C. Projection of Radar Data Cube

After the raw data is acquired, it can be reshaped into a 3D ($number\ of\ ADC\ samples \times number\ of\ chirps \times number\ of\ channels$) complex array (i.e., radar data cube). Projection of each channel vector, $\vec{x}_{ij}$, where i is the fast time and j is the slow time index, onto an angular subspace of $\mathbf{A}$ can be computed as:

$$\vec{\hat{x}}_{ij} = (\mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}})\vec{x}_{ij} \tag{4}$$

where $\mathbf{B}$ is the angular subspace defined as a subset of $\mathbf{A}$ containing the angle interval to be projected on, and $\vec{\hat{x}}_{ij}$ is the projected channel vector. A new projected radar data cube (RDC) can be constructed by repeating $\vec{\hat{x}}_{ij}$ for each i and j.

D. Micro-Doppler Spectrogram Generation

Micro-Doppler (μ-D) spectrogram is a time-frequency representation of RF data, and it can be calculated from any channel of RDC by taking the square modulus of the short-time Fourier Transform (STFT) across slow-time. A sample spectrogram for walking towards and away from radar are shown in Figure 2.

E. Similarity of Projected Spectrograms

Projection of an RDC on $\mathbf{B}$ greatly affects the resulting new RDC by introducing more emphasis on the targets located in the angle interval spanned by $\mathbf{B}$, and fading out the signatures of targets located outside the projected interval. In order to quantitatively assess the quality of the newly generated RDCs, the following steps are followed:

- **Step 1:** Generate RDC_1 and RDC_2 which are recorded from two different angles (i.e., left and right, respectively) and their corresponding μ-D spectrograms.
- **Step 2:** Merge two RDCs by adding them up to create a combined RDC, RDC_3, which contains the data of both RDCs.
- **Step 3:** Project each channel vector of RDC_3 onto $\mathbf{B}_{left}$ and $\mathbf{B}_{right}$ to obtain left and right projected RDCs, RDC_4 and RDC_5, respectively, and their μ-D spectrograms.
- **Step 4:** Compute the similarity of the spectrograms between RDC_1-RDC_4 vs. RDC_1-RDC_5, and RDC_2-RDC_4 vs. RDC_2-RDC_5.

It is expected to see that the spectrograms generated from RDC_1 and RDC_4 should have higher similarity than those RDC_1 and RDC_5. Similarly, RDC_2 and RDC_5 should have higher similarity than RDC_2 and RDC_4.

Figure 3 visually illustrates the aforementioned steps and μ-D spectrograms of each RDC. In Figure 3a, a person is walking towards and away from radar from -40° (left) and 40° (right), in RDC_1 and RDC_2, respectively. While RDC_3 is the resulting RDC after merging of RDC_1 and RDC_2, RDC_4 and RDC_5 are the RDCs generated after projection of RDC_3 onto $\mathbf{B}_{[-60^\circ\ -30^\circ]}$ and $\mathbf{B}_{[30^\circ\ 60^\circ]}$, respectively. As can be seen, while RDC_4 put more emphasis on the target approaching from left and fading out the target approaching from right, RDC_5 put more emphasis on the target approaching from right and fading out the target approaching from left. In order to verify the effectiveness of the method on the ASL data, RF recordings of a hand moving towards and away from radar from negative and positive angles are also observed. Figure 3b, shows a similar process for the hand motion data. While RDC_4 strengthens the signature of the target approaching from left and weakens the target approaching from right, RDC_5 strengthens the signature of the target approaching from right and weakens the target approaching from left. Although the separation is visually not very distinct, it can still give us information about the arrival direction of a target only by looking at the μ-D signature.

Figure 4 shows the spectrograms after projection of the words YOU, HELLO, WALK, DRINK onto the angle interval of [$\pm$30° $\pm$60°]. It can be observed that for the word HELLO, original spectrogram has almost equal signal power for each positive and negative peak. On the other hand, after projection of RDC to the left, signal power of first negative and second positive peaks increases, while those of first positive and last negative decreases. This imperfection is stemming from non-orthogonality of the steering angles.

Figure 5 shows the auto-correlation matrix $\mathbf{R}_{AA} = \mathbf{A}\mathbf{A}^T$. As can be seen, the correlation between opposing angles are not exactly 0, and this phenomenon causes leftover μ-D signatures in the projected spectrograms from the opposing angles.

Table II presents the similarity results of the spectrogram matrices of the walking data in terms of structural similarity index (SSI) and mean squared error (MSE) for three different angle intervals, namely, 0 to $\pm$30, $\pm$30 to $\pm$60, and $\pm$60 to $\pm$90. It can be observed that the rows in which the original

TABLE II: Similarity Results for the Walking Data

Original Target Angle	Projected Angle Interval	SSI	MSE
Left (-40°)	Left (-30° to 0°)	0.08	8.59e16
	Right (0° to 30°)	0.056	8.99e16
Right (40°)	Left (-30° to 0°)	0.013	5.12e19
	Right (0° to 30°)	0.1242	8.16e16
Left (-40°)	Left (-60° to -30°)	0.0088	2.56e17
	Right (30° to 60°)	0.0036	6.6e17
Right (40°)	Left (-60° to -30°)	0.0074	5.9e17
	Right (30° to 60°)	0.1301	8.41e15
Left (-40°)	Left (-90° to -60°)	0.1408	3.098e17
	Right (60° to 90°)	0.0113	1.87e19
Right (40°)	Left (-90° to -60°)	0.0026	1.53e20
	Right (60° to 90°)	0.0713	3.12e18

TABLE III: Similarity Results for the Hand Moving Data

Original Target Angle	Projected Angle Interval	SSI	MSE
Left (-40°)	Left (-30° to 0°)	0.1788	9.36e15
	Right (0° to 30°)	0.1123	6.6e15
Right (40°)	Left (-30° to 0°)	0.0117	5.13e17
	Right (0° to 30°)	0.1297	7.17e15
Left (-40°)	Left (-60° to -30°)	0.1396	0.86e16
	Right (30° to 60°)	0.0802	1.42e16
Right (40°)	Left (-60° to -30°)	0.0074	1.11e18
	Right (30° to 60°)	0.1279	1.46e16
Left (-40°)	Left (-90° to -60°)	0.2576	3.74e15
	Right (60° to 90°)	0.0293	1.45e17
Right (40°)	Left (-90° to -60°)	0.0019	3.77e18
	Right (60° to 90°)	0.0452	9.43e17

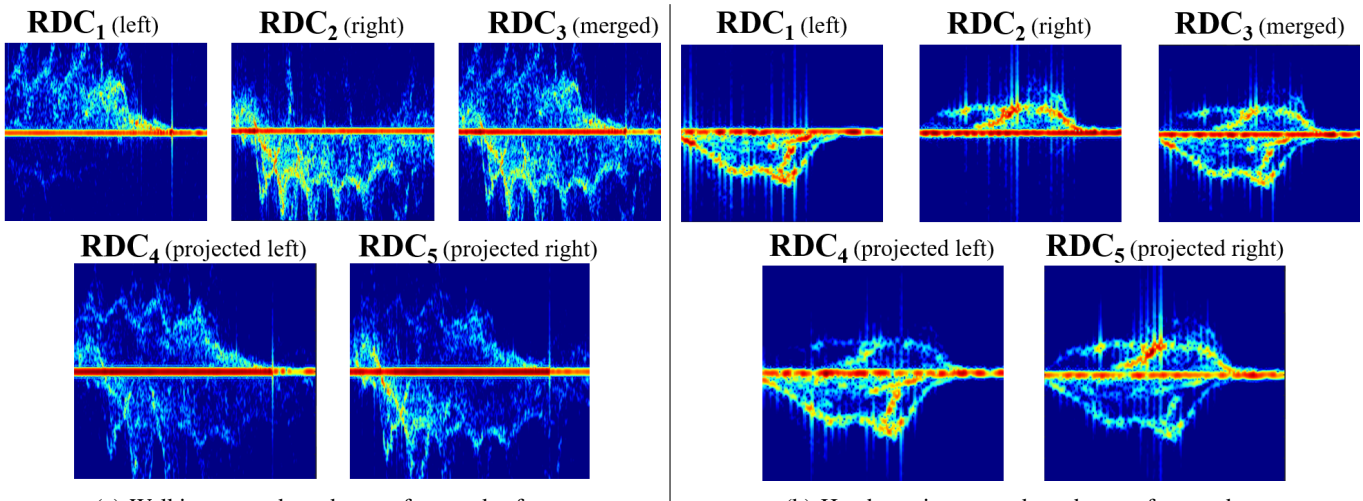
target angle and the projected angle interval matches (i.e., left to left or right to right) have higher SSI and lower MSE than the opposing projected angle interval. Similarly, Table III presents the similarity results for the hand moving data. Hand moving results are also consistent with what has been observed in the walking data except for one case where MSE between the original left and the projected left is higher than those between the original left and the projected right, which is highlighted in red in the table. This can potentially occur due to the ambiguity and spread of the target angle across positive and negative angles as it gets closer to the radar.

IV. Classification

Impact of the projection method on the classification of the ASL signs are evaluated by comparing the classification accuracies of two convolutional neural network (CNN) models. While the first model is a 4 layer single input CNN followed by fully connected layers, the second model takes left and right spectrograms as separate inputs and parallelly processes in two branches. The CNN blocks are followed by a *concatenation* layer to fuse the feature spaces of two spectrograms and encode them into a single latent space for prediction. However, using two branches doubles the number of parameters to train, and this causes over-fitting in the trained model. In order to mitigate this problem, weights across two branches are shared, hence the number of parameters to train stays the same with the baseline model.

The overall architecture of the proposed network is depicted in Figure 6. Table IV presents the classification results of 20 ASL words with and without the projection method. Each model is trained for five times to obtain consistent results. As can been seen from the table, multi-branch networks (rightmost 3 column) which take projected spectrograms as inputs outperform the baseline method with no projection.

Figure 7 presents the confusion matrix of the best performing model with **92.6%** accuracy for 20 signs. As can be seen, the words Hello and Knife have the highest error rate of 33.3%. This can potentially be due to the amount of variance in the way of signing among participants (e.g., inconsistent number of hand strokes, signing duration and speed etc.). It can also be noticed that while the signs with high radial motion (You, Well, Health, Earthquake, Breathe, Go) have the highest recognition rate of 100%, the signs with low radial motion (Lawyer, Help) and self-occlusion (Write) have



(a) Walking towards and away from radar from.

(b) Hand moving towards and away from radar.

Fig. 3: Original, merged and projected spectrograms generated from RDC_1 to RDC_5.

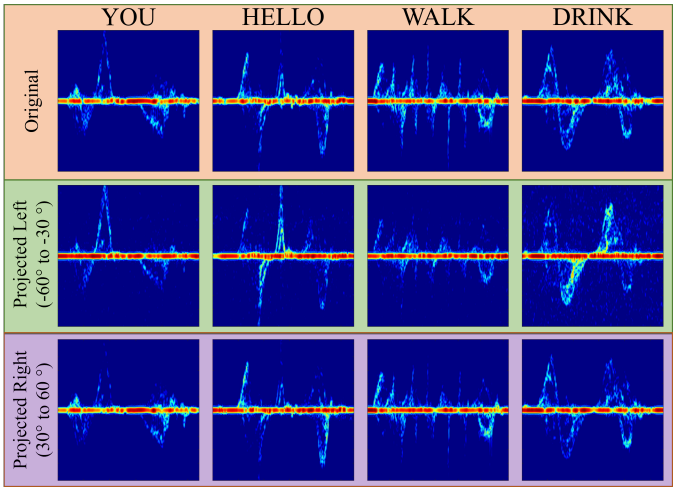


Fig. 4: Projection of the words YOU, HELLO, WALK, DRINK onto the angle interval of [±30° ±60°]

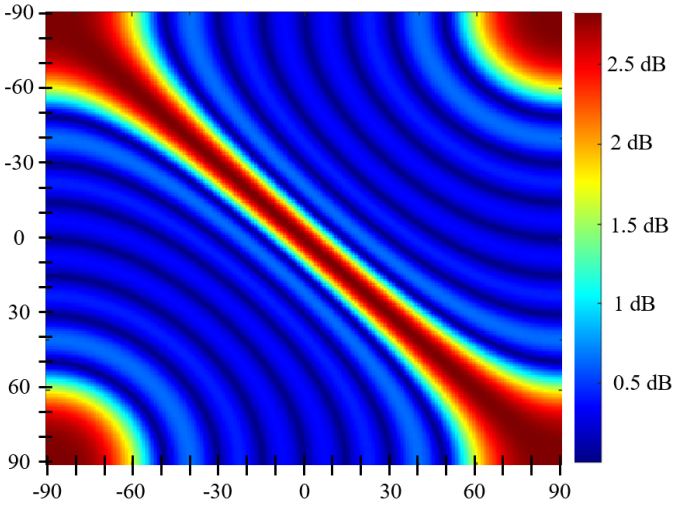


Fig. 5: Auto-correlation matrix $\mathbf{R}_{AA}$.

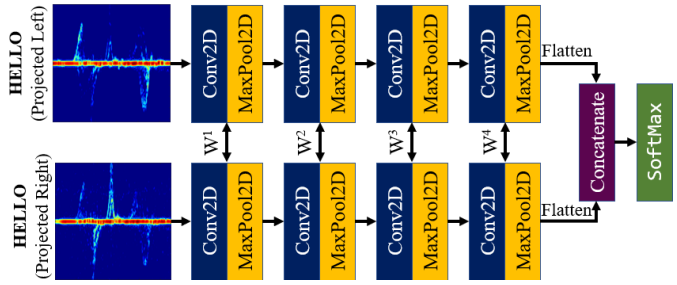


Fig. 6: Proposed multi-branch CNN model architecture where W^i denotes the weights belonging i^{th} layer of each branch.

lower recognition rates.

V. Conclusion

This work presents the initial work on projection of RF data onto a certain angle interval using a subset of the steering matrix. Although the proposed method cannot completely isolate the targets located in opposing angles, it does not require any super resolution algorithm with high computational cost. It is shown that the spectrograms generated from projected RDCs can partially recover the original RDCs before the merging operation. The quantitative results are presented in terms of SSI and MSE metrics.

A multi-branch network with shared weights across parallel layers is proposed to encode the information from positive and negative angles jointly. A mean classification accuracy of 92% is achieved for the classification of 20 ASL signs. Our future work is going to be to implement the proposed method in a multi-person scenario, and enhance the separation capability of the method, perhaps by designing more sophisticated neural networks which can extract the angular features in an online fashion.

TABLE IV: Classification Accuracy (%) Comparison Results.

Experiment ID	No Projection	Projection Angle Interval		
		0 to ±30	±30 to ±60	±60 to ±90
1	87.8	92.1	91	90.5
2	89.4	91.5	89.4	91.5
3	88.4	91.5	91.5	89.4
4	91.5	92.1	90.5	91
5	90.5	92.6	89.4	91
Mean	89.5	**92**	90.4	90.7

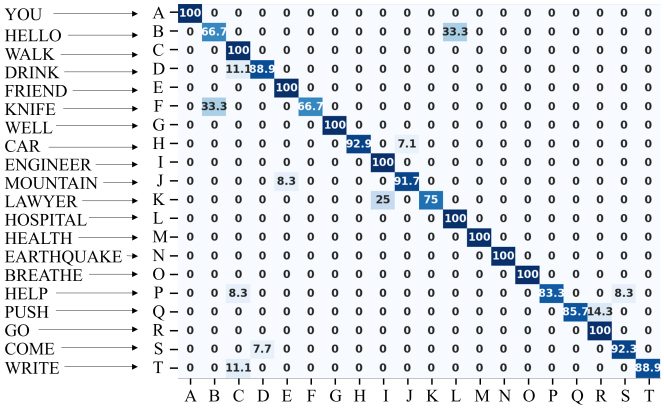


Fig. 7: Confusion matrix of the multi-branch CNN model for 20 ASL words.

Acknowledgment

This work was funded in part by the National Science Foundation (NSF) Awards #1932547, #1931861, and #1734938. Human studies research was conducted under UA Institutional Review Board (IRB) Protocol #18-06-1271.

References

[1] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 1593–1600.

[2] G.-F. He, S.-K. Kang, W.-C. Song, and S.-T. Jung, "Real-time gesture recognition using 3d depth camera," in *2011 IEEE 2nd International Conference on Software Engineering and Service Science*, 2011, pp. 187–190.

[3] Q. Yang, W. Ding, X. Zhou, D. Zhao, and S. Yan, "Leap motion hand gesture recognition based on deep neural network," in *2020 Chinese Control And Decision Conference (CCDC)*, 2020, pp. 2089–2093.

[4] K. Assaleh, T. Shanableh, M. Fanaswala, F. Amin, and H. Bajaj, "Continuous arabic sign language recognition in user dependent mode," *JILSA*, vol. 2, 01 2010.

[5] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. S. Crawford, M. M. Rahman, E. Kurtoglu, R. Aksu, T. Macks, and R. Mdrafi, "American sign language recognition using rf sensing," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3763–3775, 2021.

[6] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "Dnn transfer learning from diversified micro-doppler for motion classification," 2018.

[7] E. Kurtoglu, A. Gurbuz, E. Malaia, D. Griffin, C. Crawford, and S. Gurbuz, "Sequential classification of asl signs in the context of daily living using rf sensing," 05 2021, pp. 1–6.

[8] M. S. Seyfioğlu, A. M. Özbayoğlu, and S. Z. Gürbüz, "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 4, pp. 1709–1723, 2018.

[9] Y. Sun, T. Fei, X. Li, A. Warnecke, E. Warsitz, and N. Pohl, "Real-time radar-based gesture detection and recognition built in an edge-computing platform," *IEEE Sensors Journal*, vol. 20, no. 18, pp. 10 706–10 716, 2020.

[10] M. Amin, "Radar for indoor monitoring: Detection, classification, and assessment," 2017.

[11] W. Massagram, V. M. Lubecke, A. HØst-Madsen, and O. Boric-Lubecke, "Assessment of heart rate variability and respiratory sinus arrhythmia via doppler radar," *IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 10, pp. 2542–2549, 2009.

[12] A. Rahman, V. M. Lubecke, O. Boric–Lubecke, J. H. Prins, and T. Sakamoto, "Doppler radar techniques for accurate respiration characterization and subject identification," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 2, pp. 350–359, 2018.

[13] A. Dell'Aversano, A. Natale, A. Buonanno, and R. Solimene, "Through the wall breathing detection by means of a doppler radar and music algorithm," *IEEE Sensors Letters*, vol. 1, no. 3, pp. 1–4, 2017.

[14] A. Seifert, A. M. Zoubir, and M. G. Amin, "Radar classification of human gait abnormality based on sum-of-harmonics analysis," in *2018 IEEE Radar Conference (RadarConf18)*, 2018, pp. 0940–0945.

[15] S. Z. Gurbuz, C. Clemente, A. Balleri, and J. J. Soraghan, "Micro-doppler-based in-home aided and unaided walking recognition with multiple radar and sonar systems," *IET Radar, Sonar Navigation*, vol. 11, no. 1, pp. 107–115, 2017.

[16] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. Crawford, M. M. Rahman, R. Aksu, E. Kurtoglu, R. Mdrafi, A. Anbuselvam, T. Macks, and E. Ozcelik, "A linguistic perspective on radar micro-doppler analysis of american sign language," in *2020 IEEE International Radar Conference (RADAR)*, 2020, pp. 232–237.

[17] N. K. Caselli, Z. S. Sehyr, A. M. Cohen-Goldberg, and K. Emmorey, "Asl-lex: A lexical database of american sign language," *Behavior Research Methods*, vol. 49, no. 2, pp. 784–801, Apr 2017.

[18] V. Chen, *The Micro-Doppler Effect in Radar, Second Edition*, 2019.

[19] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 10, pp. 1797–1801, 2014.

[20] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," 2020.