Performance Comparison of Radar and Video for American Sign Language Recognition

M. Mahbubur Rahman[1], Emre Kurtoğlu[1], Muhammet Taskin[2], Kudret Esme[2],
Ali C. Gurbuz[2], Evie Malaia[3], Sevgi Z. Gurbuz[1]
[1]*Dept. of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, AL, USA*
[2]*Dept. of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA*
[3]*Dept. of Communication Disorders, The University of Alabama, Tuscaloosa, AL, USA*
mrahman17@crimson.ua.edu, szgurbuz@ua.edu, gurbuz@ece.msstate.edu

Abstract—**In the past decade, there has been a great research in the developments of American Sign Language (ASL) enabled user interfaces and smart environments, especially using wearables, RGB and RGB-D video cameras , and radio frequency (RF) sensors. Each sensor modality provides distinct advantages and suffer from various problems. Although each sensor modality is studied for ASL recognition a comparison of video and RF based sensing performance in terms of ASL recognition is not available. This study aims to compare word level ASL recognition performance over the same 100 ASL glosses data from both RF and video sensors. A top-5 accuracy of 93% was achieved while using the RF micro-Doppler spectrogram representation in a convolutional neural network (CNN) classifier, whereas with video ASL data for the same 100 words, a top-5 accuracy of 90% was achieved. This shows that radar has comparable recognition performance to video for ASL recognition.**

Index Terms—**American Sign Language, video, radar, RF sensors, deep neural networks**

I. Introduction

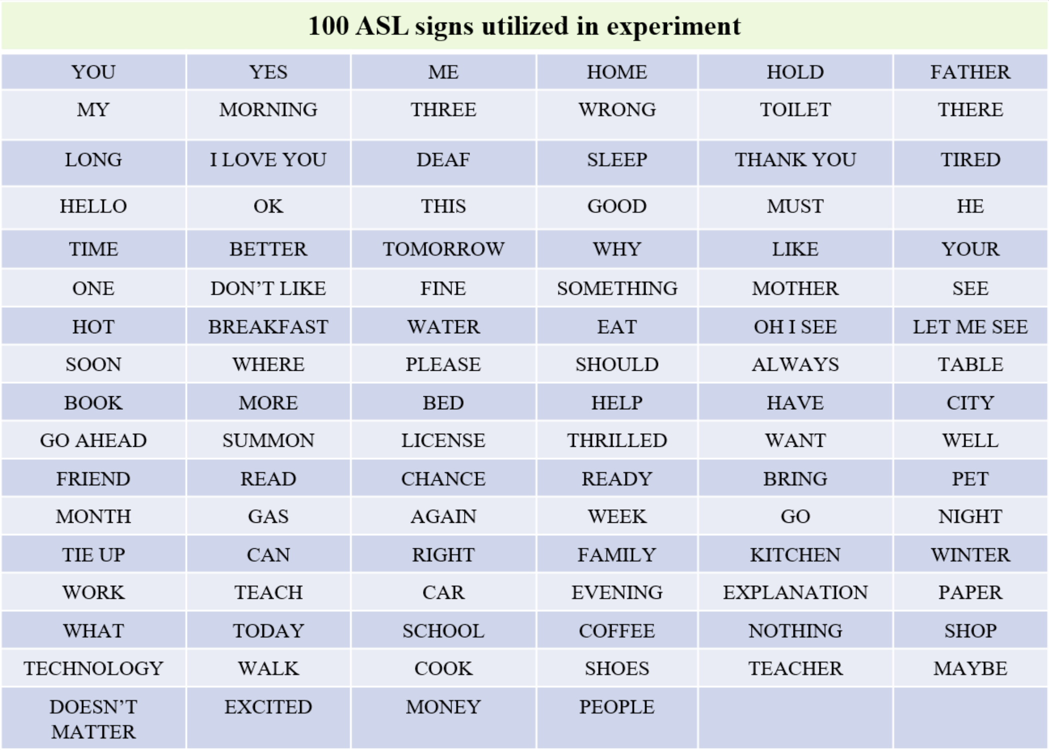
Technologies for human-computer interaction (HCI) are mostly driven by human voice, hence exclude the people from the Deaf community as they use sign language as their medium of communication. To address this, various studies have been conducted worldwide and two approaches namely, camera-based and wearable devices based sign language recognition techniques are widely adopted. The camera-based approaches [1]–[3] utilize an RGB camera and depth sensor and applies computer vision algorithms to analyze the hand gestures along with the body and facial expressions from images to recognize sign language. However, video cameras require adequate light and a direct line-of-sight to be effective. On the other hand, wearable approaches [4], [5] derive finger and hand moving patterns from multiple sensors that are attached to the user's hands or body. Although sensor-augmented wearable gloves have been reported to typically yield higher gesture recognition rates than camera-based systems, they cannot capture the information conveyed through head and body movements during signing. Also they interfere with the daily activities of a human being.

Hence, another sensing modality with the capability of non-contact sensing and work-ability in the dark, especially in a situation when camera and warbles are ineffective, is highly desirable. RF sensors can fill that void as they offer unique advantages of being non-contact, not restrictive or invasive, operate at a distance, protect the privacy of the user and personal spaces, and are effective in the dark, regardless of what the individual is wearing. In recent years, it has been shown that [6]–[10], the RF sensors can be deployed as a sign language processing technology for HCI applications. RF sensors cannot perceive hand shapes or facial expressions, but they can provide a direct measure of distance, angle and velocity as a function of time. Velocity can be obtained via the *micro-Doppler* [11] effect.The micro-Doppler signature is comprised of unique patterns directly related to the kinematics of the underlying motion, and hence capture the trajectories of hands and fingers movements during signing [12], [13], and gesturing [14], [15].

Each of the different modalities, so far discussed, have their individual advantages and limitations. None of them can be considered as the only good-to-go modality for ASL recognition. While camera and wearable systems have been compared [16], a systematic comparison of performance for RF and video based ASL recognition over a shared dataset is not available. In addition, a comprehensive approach to ASL recognition requires the integration of information from multiple data sources with different spatial and temporal scales, along with the application of linguistic knowledge about both the manual (hand motions) and the non-manual (facial expressions and body-language) aspects of ASL. A comparison of individual sensor performance also helps for developing sensor fusion techniques for this purpose.

In this work, we compare the performance of video and RF based ASL recognition over the same 100 ASL signs using state-of-the-art deep neural networks (DNNs). RF data were collected from fluent signers in a laboratory setting with an FMCW radar. The raw RF data were processed using $time - frequency$ analysis to generate micro-Doppler spectrograms for individual signs. These spectrogram images were then used to train convolutional neural networks (CNNs). However, recruiting human test subject is difficult and costly, and can result in an undue burden on Deaf participants. On the other hand, the limitation in the amount of available real training data limits the depth and accuracy of CNNs. Therefore, we

TABLE I: Listing of the 100 ASL signs utilized in experiments.



100 ASL signs utilized in experiment					
YOU	YES	ME	HOME	HOLD	FATHER
MY	MORNING	THREE	WRONG	TOILET	THERE
LONG	I LOVE YOU	DEAF	SLEEP	THANK YOU	TIRED
HELLO	OK	THIS	GOOD	MUST	HE
TIME	BETTER	TOMORROW	WHY	LIKE	YOUR
ONE	DON’T LIKE	FINE	SOMETHING	MOTHER	SEE
HOT	BREAKFAST	WATER	EAT	OH I SEE	LET ME SEE
SOON	WHERE	PLEASE	SHOULD	ALWAYS	TABLE
BOOK	MORE	BED	HELP	HAVE	CITY
GO AHEAD	SUMMON	LICENSE	THRILLED	WANT	WELL
FRIEND	READ	CHANCE	READY	BRING	PET
MONTH	GAS	AGAIN	WEEK	GO	NIGHT
TIE UP	CAN	RIGHT	FAMILY	KITCHEN	WINTER
WORK	TEACH	CAR	EVENING	EXPLANATION	PAPER
WHAT	TODAY	SCHOOL	COFFEE	NOTHING	SHOP
TECHNOLOGY	WALK	COOK	SHOES	TEACHER	MAYBE
DOESN’T MATTER	EXCITED	MONEY	PEOPLE		

proposed synthesising micro-Doppler ASL signatures with the limited real samples. A kinematically enhanced multi Discriminator branch GAN [17] architecture is proposed to generate synthetic ASL signs. Once a large datasets were generated, the Deep CNNs were trained on synthesized signatures and tested on real signer’s signatures.

For the video dataset we have extracted video samples for the same glosses as in RF case from an existing ASL video dataset; Word-LevelAmerican Sign Language (WLASL) [18]. The video samples are RGB videos with a total of 1566 video samples. Details on the extracted 100-gloss video dataset and the utilized processing is detailed in Sections II and III. Our initial results show that comparable classification accuracy levels can be achieved for both RF and video over the tested 100-glosses.

This paper is organized as follows: Section II describes the compilation of RF and video datasets for the same 100 ASL gloss. Section III describes the RF data processing and ASL recognition strategies with RF data. In section IV, video based ASL recognition methodologies are illustrated. The results of RF and video based recognition are discussed and compared in Section V. Conclusions and future research directions are discussed in Section VI.

II. RF and Video Datasets

A. RF Data Collection and Experimental Setup

The RF sensor used in this work is a TI IWR1443 77 GHz automotive short-range radar, which transmits linear frequency modulated continuous wave (FMCW) signals. The transmitted signal illuminates an ASL signer who sits 1.5 meters in front of the sensor and signs in ASL. The radar receives backscatter from the moving arms and hands, as well as reflection from static parts of the body and the environment. Thus, the signal received by the receiver is a weighted summation of time-delayed, frequency-shifted versions of the transmitted signal given by the the superposition of returns from M points on the body [19]. Thus,

$$x_{rec}(t) = \sum_{i=1}^{M} a_i exp\left\{ -j\frac{4\pi f_c}{c} R_{t,i} \right\}, \tag{1}$$

where $R_{t,i}$ is the range to the i^{th} body part at time t, f_c is the transmit center frequency, c is the speed of light, and the amplitude a_i is the square root of the power of the received signal as given by the radar range equation [20]. Thus, RF sensors provide a complex-time series of measurements in the form $x[t] = I[t] + jQ[t]$.

The data were collected in a laboratory setting, where the sensor was placed on a table at an elevation of 0.91 m from the ground. Participants sat on a chair directly facing a computer monitor, which was placed immediately behind the radar system. The monitor was used to relay prompts indicating the signs to be articulated. The radar system was positioned at a distance of 1.5 meters in front of the participant.

Four fluent ASL signers took part in the IRB-approved study, of whom 2 were Deaf and 2 were CODAs. The experiments included 100 ASL signs, as shown in Table 1, which were selected from the ASL-LEX2 [21] database to

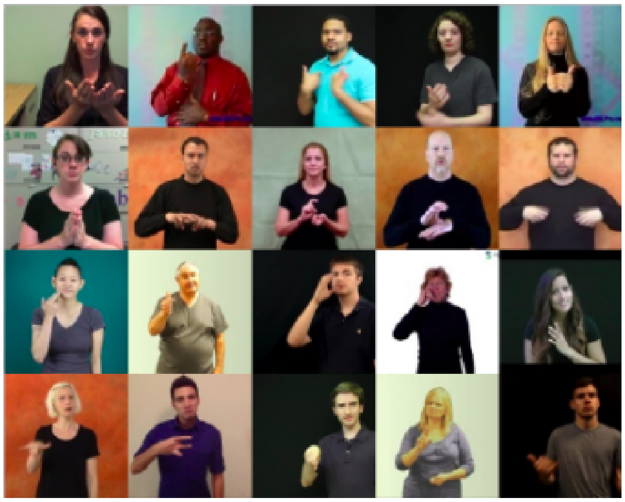


Fig. 1: Some example video frames from WLASL dataset.

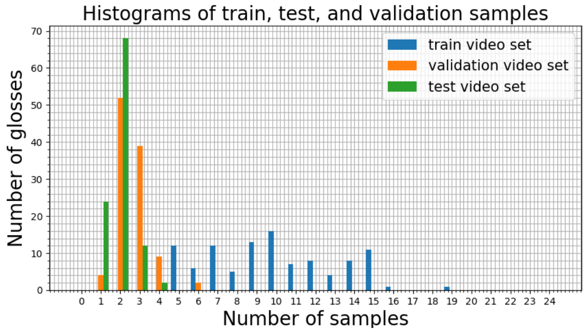


Fig. 2: Histogram of 100-gloss subset of WLASL dataset.

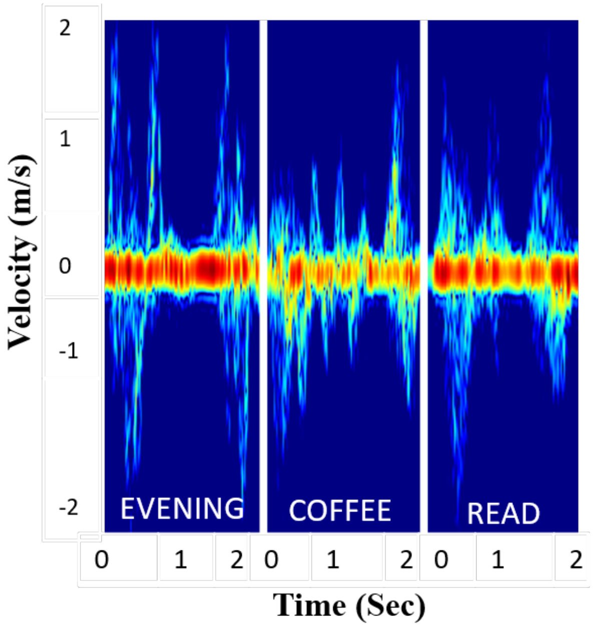


Fig. 3: Example of Micro-Doppler signatures acquired from fluent signers.

include signs of high frequency, but not phonologically related to ensure a diverse dataset in terms of both handshapes and sign kinematics. The participants repeated each sign 5 times. A total of 2000 fluent sign samples were collected over 100 ASL words [22].

B. Video Dataset

Due to the nature of the sign language problem, it is hard to collect a dataset that satisfies both quantity and quality expectations. For example, a dataset needs to contain enough videos for training, testing, and validation processes while including a variety of signers and glosses. Word-Level American Sign Language (WLASL) dataset which is introduced in [18] is an RGB video dataset that consists of 2000 different glosses including approximately 21k videos collected from 119 different signers. As can be seen from Figure 1, this dataset provides only close frontal views with different backgrounds and illuminations.

Although WLASL dataset provides a 2000 gloss dataset, in order to make a comparison with the RF data, glosses given in Table 1 are matched with the glosses provided in WLASL dataset. To the best knowledge of the authors the 100 gloss RF dataset in [22] is currently the largest RF ALS dataset. The videos corresponding to glosses provided in Table I are extracted from the WLASL dataset, this 100-gloss subset of WLASL dataset is used in our comparison analysis. Extracted video dataset consists of a total 1566 videos. As can be seen from the histogram given in Figure 2, this subset includes at least 7 videos per gloss and 15 videos in average.

III. RF Data Processing and Classification

A. Data Processing

The kinematic properties of signing results in a time-varying pattern of Micro-motions [11], e.g. rotations and vibrations, result in micro-Doppler frequencies. Each sign generates its own unique patterns, which can be revealed through time-frequency analysis. The *micro-Doppler signature*, or spectrogram, is found from the square modulus of the Short-Time Fourier Transform (STFT) of the continuous-time input signal $x(t)$ and can be expressed in terms of the window function, $h(t)$, as

$$S(t,\omega) = \left| \int_{-\infty}^{\infty} h(t-u)x(u)e^{-j\omega t}du \right|^2. \qquad (2)$$

Ground clutter from stationary objects, such as furniture and the walls, will appear in the micro-Doppler signature as a band centered around 0 Hz. At 77 GHz, elimination of low-speed signal components during clutter filtering results in performance degradation [23], therefore no filtering was applied on the data. Samples of the micro-Doppler signatures for glosses ’evening’, ’coffee’ and ’read’ are shown in Figure 3.

Compilation of large datasets for training state-of-the-art DNNs is difficult when human subjects are involved, due to the time spent in measuring numerous iterations of each class. As an initial attempt to classify the 100 ASL words, a 6 layer convolutional neural network (CNN) is utilized, and a top-1 accuracy of 56.00% were observed. The limitations in the amount of available training data limit the depth and accuracy of the DNNs utilized. Therefore, to achieve high recognition accuracy, the problem of limited training data is addressed here through generating synthetic samples from a small amount of fluent ASL signs samples using GANs.

B. Synthesis of ASL sign signatures

In general, the architecture of GANs consists of two competing neural networks i.e., generator and discriminator playing a min-max game [24].The generator network samples a predefined latent space and upsamples via transposed or deconvolutional layers to produce a synthetic image whereas the discriminator network takes that synthetic images as input and attempts to classify them as being real or fake. In our prior work [17], [25], several different types of architectures have been explored for synthetic data synthesis, including auxiliary-conditional GANs (ACGANs), conditional variational autoencoders (CVAE) and WGANs, but all were found to generate data that exhibits significant discrepancies from that of real RF signatures. While these erroneous components may not seem significant visually, they ultimately correspond to kinematically impossible articulations, which, when used as training data, incorrectly trains the DNN and significantly degrades classification accuracy.

One way to mitigate such problems is to design the GAN so as to enable greater emphasis on preservation of the shape of the envelope. The envelopes correspond to the maximum velocity towards/away from the radar; so, from the standpoint of hand kinematics, the synthetic signatures should conform to, and not exceed the envelope profiles of source data. In prior work [17], [22], [26], a multi-branch GAN (MBGAN) architecture with an additional auxiliary branch in the WGAN discriminator, which took as input the upper envelope, was proposed as a means of ensuring kinematic accuracy when synthesizing micro-Doppler signatures of different ambulatory gaits, such as walking, limping, or taking short steps. However, during production of sign language, the hands may move towards or away from the radar, so both the upper and lower envelopes are important for maintaining critical kinematic features. Hence, in this work, we incorporated two additional auxiliary branches in the discriminator: one that takes the upper envelope as input, and a second that takes the lower envelope as input. The resulting MBGAN with 3-branch discriminator is shown in Figure 4. The generator is comprised of 10 convolutional layers; each layer is followed by batch normalization with 0.9 momentum and a Rectified Linear Unit (ReLU) activation function. The main branch of the discriminator is an 8-layer CNN, where each layer is followed by a Leaky-ReLU activation function. Each auxiliary branch is comprised of three 1D-convolutional layers. The outputs of the dense layers are concatenated with the flattened output of the main discriminator. The MBGAN was trained with 80% of real spectrogram samples and a total of 50000 synthetic samples were generated over 100 ASL signs classes.

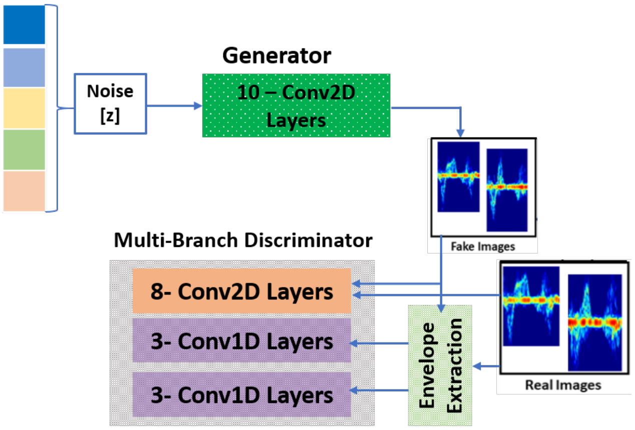


Fig. 4: Proposed 3-branch discriminator MBGAN.

IV. Video Based Processing and Classification

For sign language recognition, various different approaches have been applied in the literature such as 2D CNN with recurrent neural networks (RNN), 3D CNNs, and pose estimation based graph convolutional neural network (GCN)s [18], [27]. Each of these approaches have shown different levels of performance on video sets but as one of the highly performing approach for video based ASL recognition and as an initial attempt we developed a GCN. Without doubt, GCNs draw more attention since human skeleton can also be considered a graph. In this manner, the 100-gloss dataset drawn from WLASL was tested with two GCN architectures. The first architecture is the Temporal Graph Convolutional Network (TGCN) which is introduced in [18] combined with the human pose estimation algorithm OpenPose published in [28]. The second architecture is the Sign Language Graph Convolutional Network (SLGCN) which is introduced in [29] combined with the human pose estimation algorithm HRNet published in [30].

The OpenPose algorithm which uses a bottom-up approach provides joint information for the whole body. However, since sign language is generally about upper-body movements, not all the joints were utilized. In our test, the chosen 55 keypoints in [18] were used. These keypoints consists of 21 joints for each hands and 13 joints for other parts of upper-body. HRNet also provides whole body joint information for an image. Like in the case of OpenPose, just a portion of this joint information is used. The chosen 27 keypoints in the [29] taken into the account for this article. There were 10 keypoints for each hands and 7 keypoints for the other upper-body joints.

The GCN, which describes the architecture about the connections between the features, is constructed by using a simple adjacency matrix defining the skeleton structure. For a basic GCN, the equation can be given as,

$$x_{out} = \sigma(Ax_{in}W), \quad (3)$$

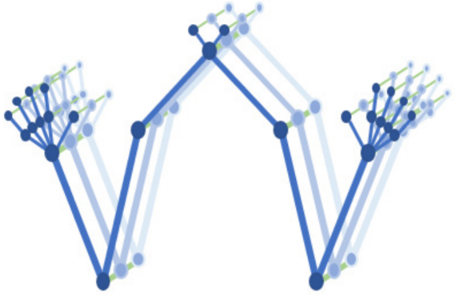


Fig. 5: Constructed graph for SLGCN architecture.

where A is the adjacency matrix, W is the trainable weight matrix, x_{in} and x_{out} are input and output features respectively. The main discrimination of two GCN methods used in this article is interpretation to graph structure of human body. While TGCN architecture considers human body as a fully connected graph, SLGCN architecture claims that a keypoint is connected with other keypoint if it is also connected in human body. The constructed graph for SLGCN is given in Figure 5. In addition to graph structure, SLGCN uses different type of features which are extracted from key points. Those extracted features are joint, bone, joint motion, and bone motion. Derivation of these features is explained in [29].

V. Results

A. RF Classification Results

The classification accuracy of 100 ASL signs for the RF data were acquired by using an deep convolutional auto-encoder (CAE) as a classifier. CAEs [31] were shown to be effective when small, yet reasonable, amounts of real data are available for training, outperforming transfer learning from weights pre-trained using ImageNet [32] for VGG [33] and Resnet [34]. Consequently, in this work, a four-blocks convolutional autoencoder (CAE) has been utilized to classify the 100-sign fluent ASL dataset. In each block, there are two convolution layers followed by a concatenation and a max-pooling layers. The filter concatenation technique concatenated a filter size of 3 × 3 and 9 × 9 to take advantage of multilevel feature extraction. CAEs use unsupervised pre-training to initialize the network near a good local minima. After training the CAE model, the decoder was removed, and two fully connected layers with 256 neurons followed by a dropout of 0.55 were added after flattening the output of the encoder. At the output, a softmax layer with 100 nodes was employed for classification. During training, an ADAM [35] optimizer was utilized, along with a batch size of 16, learning rate of 0.0005 and 30 epochs. The hyper-parameters were optimized through grid search. The classification accuracies obtained using the CAE trained on the real and synthetic data are compared in Table II with top-1,-3 and -5 cases. While using only the collected real RF dataset 56.4% average top-1 accuracy is obtained, this performance increases to 77.5% for training with the MBGAN synthesized data.

TABLE II: 100 ASL Signs Recognition Using RF Data.

Training Set	Test Set	Accuracy (%)		
		Top-1	***Top-3***	***Top-5***
Real Data	Real Data	56.35	69.12	73.54
MBGAN Synthesized	Real Data	77.51	89.26	93.00

TABLE III: 100 ASL Signs Recognition Using Video Data.

Pose Estimation Method	Architecture	Accuracy (%)		
		Top-1	***Top-5***	***Top-10***
OpenPose	TGCN	49.51	75.98	84.31
HRNet	SLGCN (Joint)	59.89	86.10	92.51
	SLGCN (Joint Motion)	25.13	50.80	66.31
	SLGCN (Bone)	53.48	81.28	93.05
	SLGCN (Bone Motion)	14.97	35.83	53.48
	SLGCN (Ensemble)	58.28	90.37	96.25

B. Video Classification Results

Table III shows the obtained results of video based data. TGCN and SLGCN results with four different features are provided. We also provide an ensemble result that combines the four features used in SLGCN. This ensemble is made by weighting results of previously given features and the weights are obtained from [29]. One can see that from the table III, the best accuracy is obtained from SLGCN which utilizes joint features with an accuracy level of 59.9%. While other features did not provide as high accuracy as bone and joint, the ensemble of all features provided an accuracy of 58.3%. Comparable classification accuracy is obtained for the developed approaches to the ones provided in [18] for a different 100-gloss dataset which consists of glosses with highest number of video samples from WLASL dataset.

VI. Conclusions

In this paper an initial attempt to compare RF and video based ASL recognition performance over the same 100-gloss set is presented. Both datasets are comparable in size with a total of 2000 RF and 1566 video data samples. It is observed that while a CNN trained on only the experimental RF data inputs such as the spectrogram images provides a 56.4% accuracy, a skeleton pose estimation based GCN with joint features provide 59.9% accuracy. Hence RF and video shows comparable top-1 accuracy results when trained on only on real data. It was also shown that if training is performed for the RF data over a MBGAN based synthetically generated dataset and tested over the real experimental data, the classification accuracy increases to 77.5%.

While this study is an initial attempt to compare RF and video based classification performance for ASL recognition, for the final version of the paper our goal is to provide results on dependence of this comparison to the number of data samples for both RF and video. In addition, we will include comparisons from similar parameter size deep learning approaches.

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] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2306–2320, 2020.

[2] C. Sun, T. Zhang, B. Bao, C. Xu, and T. Mei, "Discriminative exemplar coding for sign language recognition with kinect," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1418–1428, 2013.

[3] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, "A modified lstm model for continuous sign language recognition using leap motion," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 7056–7063, 2019.

[4] V. E. Kosmidou and L. J. Hadjileontiadis, "Sign language recognition using intrinsic-mode sample entropy on semg and accelerometer data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 12, pp. 2879–2890, 2009.

[5] Y. Li, X. Chen, X. Zhang, K. Wang, and Z. J. Wang, "A sign-component-based framework for chinese sign language recognition using accelerometer and semg data," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, pp. 2695–2704, 2012.

[6] S. Gurbuz, A. Gurbuz, C. Crawford, and D. Griffin, "Radar-based methods and apparatus for communication and interpretation of sign languages," in *U.S. Patent Application No. US2020/0334452 (Invention Disclosure filed Feb. 2018; Provisional Patent App. filed Apr. 2019.)*, October 2020.

[7] 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.

[8] S. Z. Gurbuz, M. Mahbubur Rahman, E. Kurtoglu, E. Malaia, A. C. Gurbuz, D. J. Griffin, and C. Crawford, "Multi-frequency rf sensor fusion for word-level fluent asl recognition," *IEEE Sensors Journal*, pp. 1–1, 2021.

[9] M. M. Rahman, E. Kurtoglu, A. C. Gurbuz, and S. Z. Gurbuz, "Word-level asl recognition and trigger sign detection with rf sensors," in *IEEE ICASSP*, 2021.

[10] M. M. Rahman, R. MdRafi, A. C. Gurbuz, and S. Z. Gurbuz, "Word-level sign language recognition using linguistic adaptation of 77 ghz fmcw radar data," in *IEEE Radar Conf.*, 2021.

[11] V. Chen, *The Micro-Doppler Effect in Radar, 2nd Ed.* Boston: Artech House, 2019.

[12] Sevgi Zubeyde Gurbuz, Carmine Clemente, Alessio Balleri, and John J. Soraghan, "Micro-doppler-based in-home aided and unaided walking recognition with multiple radar and sonar systems," *IET Radar, Sonar & Navigation*, vol. 11, pp. 107–115(8), January 2017.

[13] M. M. Rahman and S. Z. Gurbuz, "Multi-frequency rf sensor data adaptation for motion recognition with multi-modal deep learning," in *IEEE RADAR conf 2020*, 2021.

[14] A. Arbabian, S. Callender, S. Kang, M. Rangwala, and A. Niknejad, "A 94 GHz mm-wave-to-baseband pulsed-radar transceiver with applications in imaging and gesture recognition," *Solid-State Circuits, IEEE Journal of*, vol. 48, pp. 1055–1071, 04 2013.

[15] 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 Trans on Human-Machine Systems*, vol. 51, no. 1, pp. 32–43, 2021.

[16] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.

[17] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Synthesis of micro-doppler signatures for abnormal gait using multi-branch discriminator with embedded kinematics," in *IEEE Radar Conf.*, 2020, pp. 175–179.

[18] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1459–1469.

[19] P. van Dorp and F. Groen, "Human walking estimation with radar," *IET Radar, Sonar and Navigation*, vol. 150, pp. 356–365(9), 2003.

[20] M. Richards. McGraw-Hill Education, 2014.

[21] N. Caselli, Z. Sehyr, A. Cohen-Goldberg, and K. Emmorey, "Asl-lex: A lexical database of american sign language," *Behavior Research Methods*, vol. 49, 05 2016.

[22] M. M. Rahman, E. Malaia, A. Gurbuz, D. Griffin, C. Crawford, and S. Z. Gurbuz, "Effect of kinematics and fluency in adversarial synthetic data generation for asl recognition with rf sensors," *IEEE Transaction on Aerospace and Electronic System*, 2021.

[23] 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.

[24] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[25] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Motion classification using kinematically sifted acgan-synthesized radar micro-doppler signatures," *IEEE Trans on AES*, vol. 56, no. 4, pp. 3197–3213, 2020.

[26] M. M. Rahman, S. Z. Gurbuz, and M. G. Amin, "Physics-aware design of multi-branch gan for human rf micro-doppler signature synthesis," in *IEEE Radar Conf.*, 2021, pp. 1–6.

[27] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," 2021.

[28] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[29] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Sign language recognition via skeleton-aware multi-model ensemble," *arXiv preprint arXiv:2110.06161*, 2021.

[30] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[31] M. S. Seyfioğlu and S. Z. Gürbüz, "Deep neural network initialization methods for micro-doppler classification with low training sample support," *IGERS Letters*, vol. 14, no. 12, pp. 2462–2466, 2017.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd Int Conf on Learning Representations, ICLR 2015*, Y. Bengio and Y. LeCun, Eds.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[35] H. Sun, L. Gu, and B. Sun, "Adathm: Adaptive gradient method based on estimates of third-order moments," in *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, 2019, pp. 361–366.