Arabic Sign Language Recognition using Faster R-CNN

Rahaf Abdulaziz Alawwad[1], Ouiem Bchir[2], Mohamed Maher Ben Ismail[3]
College of Computer and Information Sciences, King Saud University, Riyadh, KSA

***Abstract*—Deafness does not restrict its negative effect on the person's hearing, but rather on all aspect of their daily life. Moreover, hearing people aggravated the issue through their reluctance to learn sign language. This resulted in a constant need for human translators to assist deaf person which represents a real obstacle for their social life. Therefore, automatic sign language translation emerged as an urgent need for the community. The availability and the widespread use of mobile phones equipped with digital cameras promoted the design of image-based Arabic Sign Language (ArSL) recognition systems. In this work, we introduce a new ArSL recognition system that is able to localize and recognize the alphabet of the Arabic sign language using a Faster Region-based Convolutional Neural Network (R-CNN). Specifically, faster R-CNN is designed to extract and map the image features, and learn the position of the hand in a given image. Additionally, the proposed approach alleviates both challenges; the choice of the relevant features used to encode the sign visual descriptors, and the segmentation task intended to determine the hand region. For the implementation and the assessment of the proposed Faster R-CNN based sign recognition system, we exploited VGG-16 and ResNet-18 models, and we collected a real ArSL image dataset. The proposed approach yielded 93% accuracy and confirmed the robustness of the proposed model against drastic background variations in the captured scenes.**

Keywords*—Arabic sign language recognition; supervised learning; deep learning; faster region based convolutional neural network

I. Introduction

Gesturing is one of the earliest forms of human communication. Nowadays, Deaf and Hard of Hearing (DHH) people are the predominant users of the officially recognized sign language which consists of alphabets, numbers, and words typically used to communicate within and outside their community. Typically, a sign language consists of; (i) manual components, and (ii) non-manual component. Specifically, the configuration, the position, and the movement of the hands form the manual components. On the other hand, the facial expression and the body movement compose the non-manual components. Such sign language is perceived as a non-verbal communication way that is mainly intended to ease the communication for the DHH persons. However, the communication between a Deaf person and a hearing individual remains an open challenge for the community. In fact, approximately 466 million people who suffer from a moderate to profound hearing loss struggle with communication daily. In other words, deaf people cannot be considered as a linguistic minority which the language can be neglected.

A sign language includes designated hand gestures for each letter of the alphabet. These gestures are used to spell people names, places, and other words without a predefined sign. Besides, it is a common occurrence for the sign formation to resemble the shape of the written letter. Although the hand gestures exhibit some similarities due to the limited number of possible hand gestures, sign language is not universal. Specifically, there are 144 sign languages around the world [44]. They vary based on the region/country rather than the language itself. For instance, The Arabic Sign Language (ArSL) includes 30 identical alphabet signs. Fig. 1 shows the sign corresponding to the letter "V" in the British and American sign languages respectively.

Despite the variations noticed on the same sign gesture when performed by signers from different origins and/or having different background, the discrepancy remains minor and affects few letters only. Particularly, the "Ra" and "H" letters can be expressed either dynamically or statically depending on the signer preference. Also, the letter "Jeem" which is represented using a curved palm, can be performed using either a sharp or a soft palm. In order to overcome such discrepancies, a considerable effort was made to unify ArSL and come up with a standard language that can be understood and used by all Arab DHH [1]. Nevertheless, fingerspelling can still be used as a common and standard way of communication between Deaf Arabs.

The semantic meaning of the gesture is a main property of the ArSL. For example, the pointing finger in the three letters "Ba", "Ta", and "Tha" represents the number of dots that the letter has. Moreover, ArSL has the specificity of having similarities within the sign language alphabet. For instance, as depicted in Fig. 2, the letter pairs "T'a" and "Th'a", "Ayn" and "Ghayn", and "Dal" and "Thal" exhibit highly similar visual properties. This makes the recognition task even more challenging for these letters.

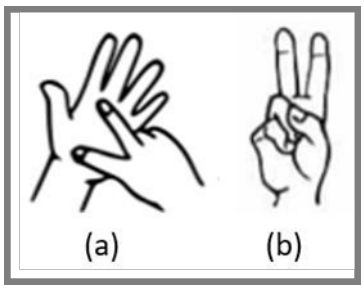

Fig. 1. Sign of Letter "V" (a) British Sign Language, (b) American Sign Language.

Fig. 2. Unified ArSL Alphabet.

Deafness can be a social barrier especially due to the hearing people's reluctance to learn a new language exclusively mastered and used by a minority. In fact, this unwillingness takes deaf persons to a state of isolation and detachment. However, the recent technological advances have promoted the development of sign language recognition systems [2-4] for different sign languages, such as Chinese Sign Language, British Sign Language, American Sign Language. One should mention that no such contributions have been achieved for the uniform Arabic Sign Language (ArSL) recognition due to the discrepancies between speakers from different Arab countries [5]. Despite this inconsistency at the language level, the hand gestures of the ArSL letters and numbers are identical for all DHH Arabs.

ArSL had its share of sensor-based systems, which the usability was mainly affected by the mandatory use of gadgets such as gloves. In other words, such solutions are intrusive and suffer from a lack of usability. Lately, image-based systems have alleviated this problem and provided a non-cumbersome solution where signs are translated using smart device cameras. Ideally, a real-time Arabic sign language recognition system would assist DHH persons and reduce their constant dependence on human translators. In particular, it would help them interact and socialize better with hearing persons. Typically, image-based solutions rely on image processing to segment the hand region, and use machine learning techniques to map the captured gestures into the pre-defined letter classes. Specifically, the image is conveyed as input, and the hand is then segmented to separate it from the background. Next, the obtained object is provided as input to the machine learning model. Note that to segment the hand, appropriate features need to be extracted from the image. These features are intended to ease the discrimination between the hand and its background. Similarly, in order to recognize the gesture, other features are used to differentiate between the different gestures classes. The choice of the appropriate features is not straightforward. It constitutes an issue for these image-based systems [4]. Moreover, the overall system performance depends on the accuracy of the segmentation task which consists of isolating the hand region from the remaining image content. In particular, the high variability of the image visual properties, as well as the similarity between the hand and the face skin color, make the segmentation even more acute.

In order choose the relevant visual descriptors and enhance the segmentation accuracy, we propose to design and implement a novel Arabic Sign Language recognition system based on the Faster Region Convolutional neural network (R-CNN). Actually, the Convolutional Neural Network (CNN) [6] is a deep learning based approach classically used for image classification [7]. The considerable learning ability of CNN is attributed to the multi-stage and hierarchal features extraction achieved by the network. The proposed CNN based approach can be perceived as an alternative to the manual feature extraction and selection needed for the segmentation and the sign recognition tasks. Furthermore, we exploit Faster Region Convolutional Neural Network (R-CNN) which performs both real-time object detection and classification to address the ArSL recognition problem.

II. Related Works

Image-based Arabic Sign Language recognition systems have tackled critical technical challenges such as the hand segmentation and the choice of the visual descriptors. On the other hand, issues such as the visual similarity between the signs of some letters like “Ra” and “Za” are specific to the Arabic sign language. Several approaches have been reported in the literature to tackle the Arabic sign language recognition [8]. Some of them extract specific features from the image and feed them into a machine learning algorithm. In the following, we refer to such solutions as conventional approaches as opposed to the latest ones based on deep learning.

The Arabic Sign language recognition system introduced in [9] coverts the input images to the YCbCr space in order to detect the hands and the face using the skin profile. A morphological operation [12] is then performed on the converted image to fill the gaps in the obtained regions. To extract features that are able to distinguish between similar signs, the Prewitt operator [10] was used to encode the edges of the hand region. Next, the Principle Component Analysis (PCA) [11] was deployed on the extracted features to reduce the dimensionality and determine the final feature. Besides, the classification task was performed using the K-Nearest Neighbor (KNN) [13] which yielded an accuracy of 97%. In [14], an ArSL finger spelling recognition system which relies on the SVM classifier [15] was proposed. The sign image was captured using a sensor that captures the image intensity and depth. The closest object to the sensor was assumed to be the signer. Another skin segmentation step is added for a better performance under complex background situations. Two features are then extracted from the segmented image. Namely, the Principle Component Analysis (PCA) [11], and the Histogram of Oriented Gradients (HOG) [16] were associated with the PCA to encode the visual properties of the image regions. The classification task was achieved using a multiclass Support Vector Machine (SVM) [15]. This yielded an outperformance of HOG-PCA due to its ability to discriminate between similar signs in addition to its robustness to local illumination

variation. Specifically, the accuracy reached 99.2% while PCA's performance attained 96.38%. In [17], the sign image is converted into the YCbCr color space for a more accurate hand segmentation. Besides, the contrast, the correlation, the energy, and the Local Homogeneity are computed from the Grey Level Co-occurrence Matrix (GLCM) [16]. The extracted feature is then fed into the Multiple Layer Perceptron [18] for skin detection. For the gesture recognition, both the outer and the inner edges are detected, and the Tchebichef [16] and Hue moments [19] are extracted. In addition, the computation of the relative area and the minimum and maximum relative distances were measured. The resulting features are then conveyed to an SVM [15] and a KNN [13] classifiers to map the input into the pre-defined classes. The proposed system was evaluated using our two ArSL datasets that include the 30 sign gestures. The first dataset was collected by 24 signers and a solid background was used for all captured scenes, while the second one which exhibits complex background was collected by 8 signers. The obtained results proved that KNN outperforms SVM with 94.67% accuracy for the first dataset and 89.35% accuracy for the second one. Similarly, the researchers in [20] compared two finger spelling recognition systems. The first one relies on KNN [13] as classifier while the second system uses the Multiple Layer Perceptron (MLP) [18] to categorize the sign images. The captured images include solid background. The signs were grouped into three categories based on the wrist orientation. One should mention that the matching operation of each sign was performed only within its allocated group. The authors introduced an edge feature to calculate the pairwise distances between the wrist and fifty equidistant contour points. The nearest neighbor and MLP were used for classification resulting in an accuracy of 91.3% and 83.7% respectively. Whereas the researchers in [22] proposed an ArSL recognition system based on the Scale-Invariant Features Transform (SIFT) [21]. Their algorithm can be summarized as: *(1)* convolve the image with Gaussian filter of different widths to create the difference of Gaussian function pyramid between filtered images, *(2)* Find the extrema in the Gaussian pyramids by comparing each point with its 26 neighbors, *(3)* Eliminate extrema key points that were suspected to be sensitive to noise or were located on an edge, *(4)* Assign orientation by forming a histogram from the gradient orientations of sample points within a region around the extrema points, and finally, *(5)* Create a descriptor for the local image region that is highly distinctive at each candidate. The dimensionality of obtained feature vector is then reduced using the Linear Discriminant Analysis (LDA) [23]. The reduced feature vector is fed to three different classifiers. Namely, the Support Vector Machine (SVM) [15], the one nearest neighbor, and the K-Nearest Neighbor (KNN) [13] were used to classify the input vectors. The results showed that SVM outperforms KNN with an accuracy of 98.9%.

In [24], an Adaptive Neuro-fuzzy Inference System (ANFIS) [25] intended to recognize the 30 alphabets of Arabic sign language was outlined. The input image was filtered using a median filter in order to reduce the noise and enhance the image for the segmentation. The latter is done using an iterative thresholding algorithm [16]. The architecture of ANFIS consists of five layers where the gesture is provided as input and the output layer indicates to the degree to what the input satisfies the rule. The overall recognition system confirmed its robustness and invariance to size, position, and direction of the input sign. However, similar gestures such as "Dal" and "Thal" were misclassified which resulted in 93.5% accuracy. Lately, the authors in [26] used two different neural networks and four visual descriptors to address the sign language recognition problem. In particular, they used the 30 letters ArSL dataset in [24] in which all images have a solid background, and the hand is the only object within the image. As a preprocessing step, the image was filtered with a Canny edge detector [27]. Specifically, the four visual descriptors used in their work were the Hu Moments [19], the Local Binary Pattern [28], the Zernike Moments [29], and the Generic Fourier Descriptor [16]. These features were provided as input to two different neural networks: MLP and Probabilistic Neural Network (PNN) [30]. The descriptors were first tested individually, then various combinations were evaluated for three different datasets. The Local Binary Pattern (LBP) descriptor yielded 90.41% accuracy when associated with PNN classifier, and it attained 86.46% accuracy when combined with MLP. Similarly, in [31], the researchers considered five features to assess their ArSL recognition system performance. Namely, they compared the Histogram of Oriented Gradients (HOG), the Edge Histogram Descriptor (EHD), the Local Binary Pattern (LBP), the Gray-Level Co-occurrence Matrix (GLCM), and the Discrete Wavelet Texture Descriptor (DWT) [16]. The descriptors were extracted from the ArSL alphabet images and classified using a One versus All SVM classifier. Their dataset by 30 was collected by 30 different signers. It includes 30 static Arabic letters with a solid background captured using a phone camera and. The obtained experiments showed that the HOG descriptor overtakes the other descriptors with an accuracy of 63.5%.

In addition to the conventional approaches, existing Arabic sign language recognition systems rely on deep learning paradigms which the ability to learn the most relevant features was confirmed in a wide range of applications. In particular, the authors in [32] designed an ArSL alphabet and digits recognition system using convolutional neural networks. Their network inspired by LeNet-5 [6] is composed of two convolutional and Leaky ReLU layers, two Max pooling layer to reduce the image size, one 75% dropout layer to reduce overfitting, and three fully connected layers for classification. The network was trained using Adam Optimizer with a learning rate of 0.03. Different ratios of training data were tested and 80% gave the best results. The evaluation was made using a collection of 5839 images for the 28 letters of ArSL and 2030 images of the decimal digits. All images include a solid background which allowed the researchers to omit the segmentation step. The experiments results showed that the proposed system outperforms other systems, and attained an accuracy of 90.02%. Similarly, a deep Recurrent Neural Network (RNN) [33] was adopted in [34] to address the Arabic sign language recognition challenge. A collection of 30 ArSL alphabets images was collected by two signers with 15 repetitions. The signers had to wear a colored glove to allow the system capture the signs. The RGB images were converted into the Hue-Saturation-intensity Value (HSV) space [16]. Then, a Fuzzy C-mean (FCM) clustering algorithm [38] was deployed to segment the different fingers. Thirty features were extracted

from the fingertips positions and orientation. In addition, four neural networks were investigated, namely, the feedforward neural network [35], the Elman neural network [36], the Jordan neural network [37], and the fully connected Recurrent Neural Network (RNN) [33]. RNN outperformed the other networks with an accuracy of 95.1%, although, the letter "Ghayn" was highly misclassified.

In [39], a deep learning recognition architecture called PCANet was introduced. Taking as input the depth image, the hand is segmented by assuming it is the closest object to the sensor. Both the RGB component and the depth component were fed individually to two different PCANet networks to automatically extract the features. PCA [11] was also deployed at the convolutional layer to find the orthogonal filters from the local patches of the input images. The learned feature vectors were next conveyed to the SVM classifier [15]. The experiments showed that the depth component achieved a better performance than the intensity component with an accuracy of 99.5%. This can be attributed to the fact that the RGB component is affected by the lighting variations and cluttered backgrounds.

The conventional and deep learning based Arabic Sign language recognition approaches reported above show that the hand segmentation is typically the first step of any sign language recognition system. The hand segmentation is a challenging task due to the difficulty to adapt to all images which exhibit highly variant levels of illumination, background complexity, skin tones and shapes. ArSL recognition systems that have been reported in the literature tackled the problem using different ways. Some works [20][22][31][26] bypassed the segmentation stage by restricting the input images to have a uniform background resulting in easier extraction of hand shape. Other approaches opted to use external equipments to aid correct capturing of the hand gesture, such as in [14][39], a Kinect sensor that captures the intensity and the depth of the images was employed. In this case, the hand is segmented as the nearest object to the camera. Similarly, in [34] a colored glove indicating the five fingertips and the wrist was used in order to recognize the signer gesture. However, the approaches in [34][14][39] imposed an unrealistic restriction to sign language recognition systems due to the inconveniency of using expensive sensors or colored gloves. On the other hand, others proposed segmentation techniques relying on skin pixel's detection as in [9][17][22]. The skin segmentation alleviates the previously mentioned problems by detecting the hand from an RGB image which does not have a uniform background without the use of any accessory or expensive sensors.

Determining the appropriate visual descriptors allows the segregation between the hand pixels and the background pixels remains an open problem. Another problem faced by ArSL recognition systems is the unavailability of large benchmark data sets with non-uniform backgrounds. In fact, small datasets such as those in [9][20] would lead to unintentional overfitting during the model learning phase. In other words, evaluating the model using small datasets may not reflect the real recognition performance. Additionally, the choice of the most suitable feature to describe the gesture can be achieved using deep learning as reported in [34][39][32]. However, to the best of our knowledge, only three works adopted deep learning to overcome the ArSL recognition challenge. All of them bypass the segmentation task by either using solid background, accessories, or depth sensors.

In this research, we propose a novel Faster R-CNN based recognition of the thirty letters of the Arabic Sign Language. The trained network is intended to segment the hand and recognize the sign gestures.

III. Proposed Method

In this research, we aim to recognize the hand gestures of the Arabic sign language using two-dimensional images, and translate them into text. The proposed system is intended to support non-hearing people in their communication with others either they master or not the ArSL language. This would lessen the social hardship this community withstands daily. Moreover, the proposed system is not a bothersome for the user since it does not require any accessory or sophisticated sensors or cameras. Specifically, we propose a faster R-CNN based approach to localize and classify the thirty letters of the Arabic sign language. In particular, a deep learning network that is designed as a typical CNN architecture is utilized as a feature extractor. The rationale behind the choice of the proposed Region CNN (R-CNN) is its noticeable impact on the object recognition field. In fact, the region proposals generation using an intelligent selective search yields to relax the need for a separate image segmentation stage. Nevertheless, some limitations were noticeable concerning the efficiency of the method, more specifically, the large number of proposals that are conveyed to the network represents a major drawback. Therefore, the more recent version fast R-CNN [40] was introduced to enhance the performance by integrating a Region of Interest (ROI) pooling layer and thus reducing the processing time required by the network. Despite this enhancement, the main issue still persists, laying within the time-consuming selective search used for proposal generation. Consequently, the latest incarnation of region CNN, namely the faster RCNN [40], was considered adapted in this research to exploit the Region Proposal Network (RPN) originally designed for real-time object recognition as depicted in Fig. 3.

The architecture of the proposed network is illustrated in Fig. 4. As it can be seen, the CNN network is utilized as feature extractor through the processing of the input image using the convolutional layers designed to produce a feature map. The Region Proposal Network (RPN) slides a window over the obtained feature maps while calculating the objectness score and the bounding box coordinates for each object (gesture) in order to produce several candidate object/regions. Lastly, given these candidate regions, the sign gesture classification task is performed by the detection network which is composed of fully connected layers.

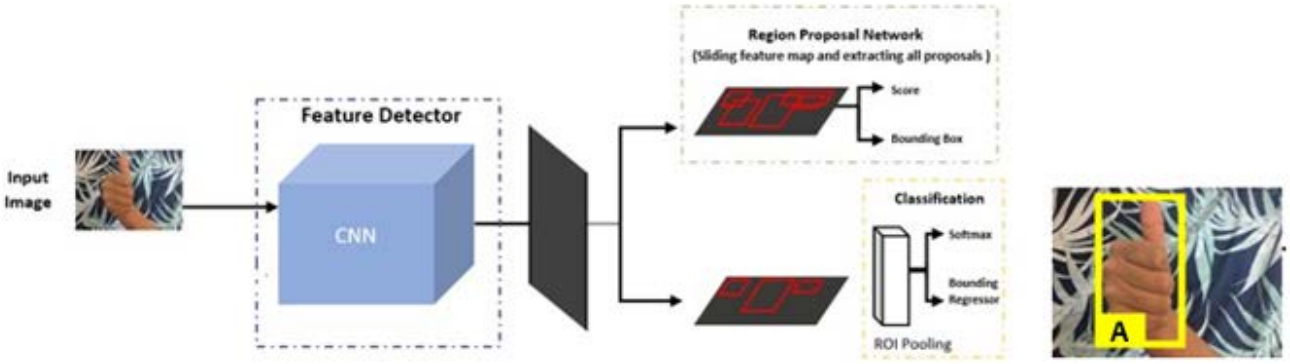

Fig. 3. The Network Architecture of the Proposed Approach.

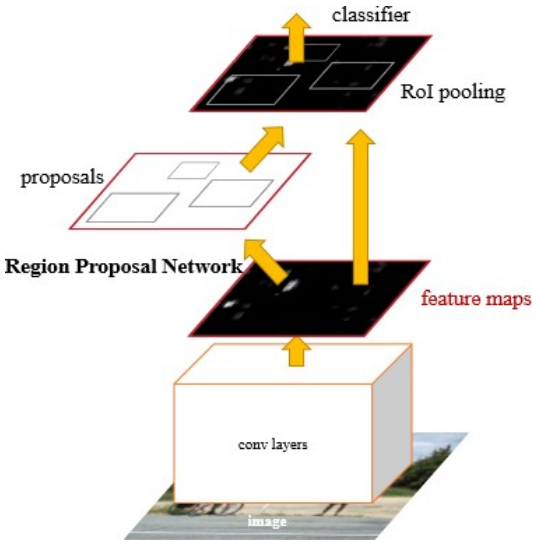

Fig. 4. Faster R-CNN Architecture.

The Region Proposal Network (RPN) in Fig. 3 can be perceived as a small pre-trained network that consists of one main convolutional layer with a 3 × 3 receptive field. It receives the feature map as input and outputs a specified number of potential region proposals. This network requires a hyperparameter k which indicates the number of rectangle boxes (anchors) of diverse ratios and scales, thereby, addressing the issue of different possible sizes of an object within the image. The initial state for each anchor is negative and it is only set to positive if the Intersect over Union (IoU) with respect to the ground truth is larger than a specified threshold. Furthermore, to contain the number of generated proposals, a Non-Maximum Suppression (NMS) is employed to remove proposals that overlap with other high score proposals. The top regions obtained using NMS are fed into the ROI layer where each region proposal includes the object scores, indicating whether the anchor containing an object, and four coordinates for the bounding box centroid as [x, y] and the width and height of the bounding box.

In this work, we investigate two architectures; The first one associates the deep VGG-16 model [41] to the faster R-CNN [40], while the second architecture relies on the ResNet architecture [42] which proved to have a faster and more accurate recognition in the ImageNet contest. These two architectures along with the pre-trained models are meant to be trained using our own ArSL gesture images. For this purpose, the top dense layer is replaced by a 1x1x31 layer that indicates the 30 classes of ArSL letters and one class for non-gesture objects.

For the training phase, the RGB image collection of the sign gestures are resized to 224 × 224 and fed to the network. The weights of the original Faster R-CNN are the starting point for our ArSL recognition network. However, the fully connected layers including the Softmax classifier and the regression box estimator are initialized from two zero-mean Gaussian distributions with a standard deviation of and 0.001 respectively. The captured images are conveyed to the feature extraction network to generate the feature maps. These maps are fed to the Region Proposal Network (RPN) in order to generate potential hand gestures. The output of the RPN contains the coordinates of the bounding box and a score indicating the existence or absence of a hand. The proposals generated by RPN are conveyed to the ROI pooling layer alongside the feature map generated by the feature extraction network. The scaled feature maps, including both bounding box and score, are fed to a fully connected layer for classification.

IV. Results

In order to conduct a comprehensive evaluation of the proposed approach, a dataset with non-uniform background and no color restrictions was collected using non-depth cameras. Specifically, our dataset includes RGB images of ArSL gestures captured using mobile cameras from both deaf and hearing signers with different hand sizes and skin tones. One should note that the existing datasets do not comply with these conditions. Fig. 5 shows sample images that correspond to the letter “Ghayn” sign. Different signers, from different nationalities, sex, and age group, performed the thirty ArSL sign gestures in various backgrounds and illumination and variation according to their sign preference. This resulted in a collection of 15,360 images of size 720 x 960 x 3. The ground truth for each image consists in the label of the gesture which is the corresponding alphabet, and the coordinates of the upper left corner (x,y) and the (width, height) of the bounding box that tightly engulfs the hand gesture. Both, the labels and the bounding box coordinates are provided and used in the learning process.

Fig. 5. Sample Images the Letter "Ghayn" from our Collected ArSL Data.

To evaluate the performance of the proposed approach to recognize each ArSL class, four standard performance measures were adopted, namely, the accuracy, the precision, the recall and the F-1 measure were used in our experiments. Note that although the detection of the hand is a critical task achieved by the proposed approach, the ultimate purpose remains the gesture recognition. Therefore, a clear focus is made on the overall recognition performance to assess the obtained results.

The models considered in this research were trained on the collected ArSL dataset. Specifically, the dataset was first split into three parts: 12,240 images (60 %) were used for training. On the other hand, 20% of the image collection was dedicated for a 3-fold cross-validation. Finally, 3060 images (20%) were reserved for testing. The resulting subsets were used to train both the VGG-16 and ResNet based networks. In order to conduct a fair comparison, we secured a uniform hyperparameter setting for VGG-16 and ResNet-18. Particularly, the starting learning rate is set to 1e-3 with a Stochastic Gradient Descent (SGD) optimizer of 0.9 momentum and a minibatch size of 1. Since a high number of epochs may lead to an accidental overfitting, a zero-patience stopping criteria was adopted in our experiments. This technique reduces the overfitting risk and provides an insight on the recognition progress during the training phase. In other words, the validation accuracy is monitored after each epoch, and at first sign of degradation the training is set to halt.

For a more objective assessment, a 3-folds cross-validation was adopted for validation in our experiments. Each fold contains 8160 images for training and 4080 images for testing. Besides, the anchor box hyperparameter, which is a critical factor for the recognition performance, was evaluated using all training images and their corresponding bounding boxes in order to find the optimal value that yields the highest IoU. Empirically, setting the number of anchor boxes to 9 yielded the best performance. Table I reports the results obtained using the two considered models; VGG-16 and ResNet-18. As it can be seen, both models yield a good performance with an accuracy around 93% with a slight edge for ResNet-18.

Although the results for both models reflect an extremely close performance, in term of training time, ResNet outperforms VGG-16. In fact, ResNet achieved its highest performance after 371 epochs while VGG-16 achieved it after 516 epochs. In order to investigate further the two models performances, we analyzed their recognition results with respect to each class. Particularly, Fig. 6 and Fig. 7 report the confusion matrix, and the performance measures obtained using the VGG16 and ResNet respectively.

As it can be seen, simple gestures like “Alef” and “Lam”, are recognized correctly despite the intra-class variation noticed in the dataset as illustrated in Fig. 7. Moreover, the two classes “Dhad” and “Ya” that exhibit similar gestures have a total of three misclassified instances only. However, similar letters like “Ra” and “Za” have relatively lower recognition rate of 86% and 83% respectively. Another letter with a low performance was “Ghaf” with an average of 83%. This is due to high similarity between “Ghaf” and the letter “TM”, despite the fact that the latter had good average recognition of 93%.

On the other hand, as shown in Fig. 8, ResNet is able to distinguish between similar letters like “Sheen” and “Seen” with only one misclassified instance. However, letters like “Ayn” and “Ghayn”, although having a high recognition rate of 90%, the 10% misclassified instances were classified as unsimilar letters. In fact, few instances of the letter “Ayn” are classified as “Jeem” and “Thal” by both models. This can be attributed to the high variance of these “Ayn” instances. The lowest recognition rate obtained by ResNet model is for the class “Za” with a value of 84%. This due to the high visual similarity between the two letters “Za” with “Ra”.

TABLE I. PERFORMANCE MEASURES OBTAINED USING VGG-16 AND RESNET-18

	Validation				***Testing***			
	Accuracy	***Precision***	***Recall***	***F1***	***Accuracy***	***Precision***	***Recall***	***F1***
ResNet-18	98.6%	98.5%	98.6%	98.5%	93.4%	93.3%	94.3%	93.7%
VGG-16	97%	97%	97%	97%	93.2%	93.6%	93.5%	93.5%

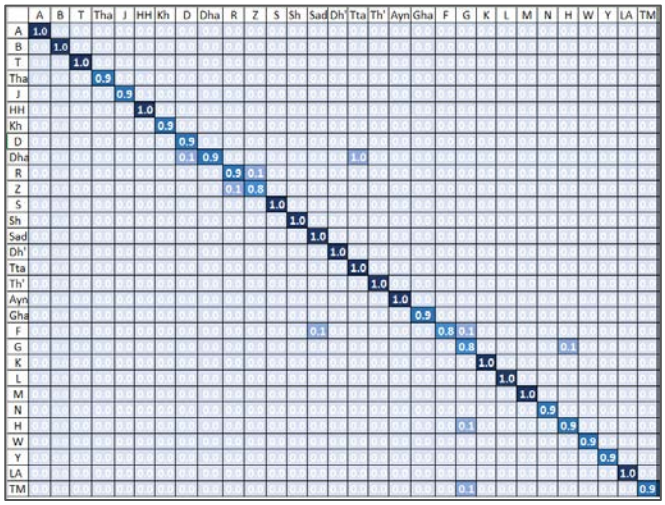

Fig. 6. Confusion Matrix for VGG-16 Obtained using the Test Set.

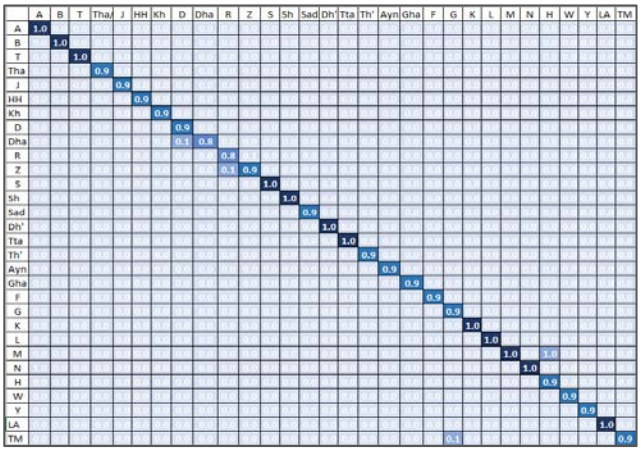

Fig. 7. Confusion Matrix for ResNet-18 Obtained using the Test Set.

Sample results for four sign gestures obtained using different models are displayed in Fig. 9. As it can be seen, in the reported result images, the sign gestures are contoured with the bounding box along with the associated confidence. One can notice that although the bounding box does not fit exactly the hand sometimes, the recognition confidences are still high.

Moreover, we compared the proposed recognition system to the most relevant state-of-the-art works that reported the highest accuracy for ArSL recognition using non-uniform background images. Specifically, we compared the results obtained by VGG-16 and ResNet to two nearest neighbor classifiers proposed in [9][17] which are based on Skin Profiling and MLP skin segmentation respectively. Table II depicts the performance comparison between the KNN based approaches [9][17], and the two Faster R-CNN approaches based on VGG-16 and ResNet respectively. The obtained results show a huge gap between the proposed Faster R-CNN approaches and the existing work in [9]. In fact, the work in [9] achieved a low detection accuracy of only 4% when implemented with the dataset we collected.

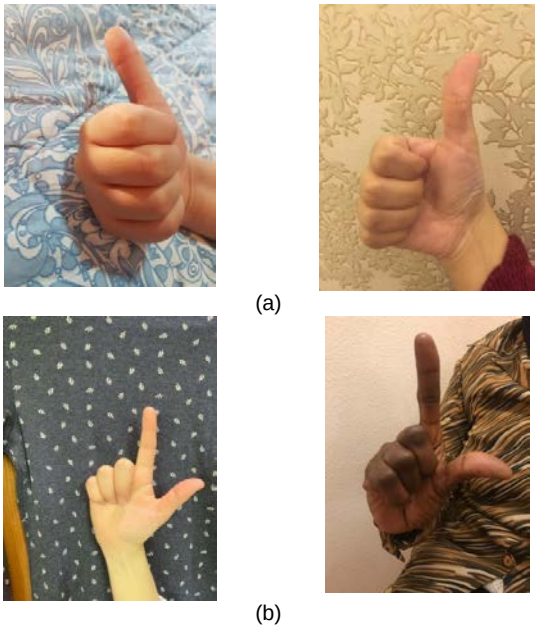

Fig. 8. Dissimilarities between the Letters: (a) “Alef”, and (b) “Lam”.

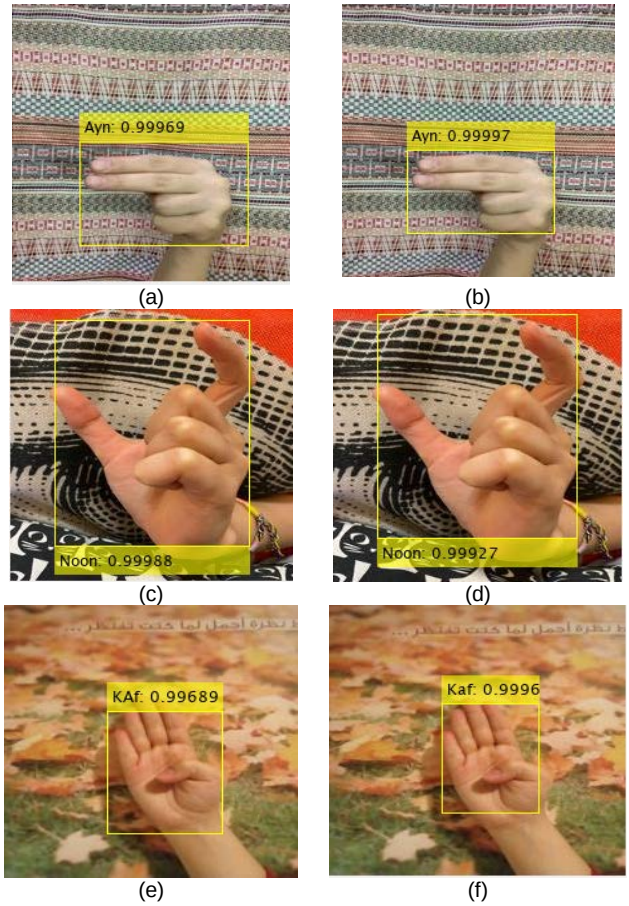

Fig. 9. Sample Recognition Results Obtained using ResNet-18 and VGG-16 for the Letters “Ayn”, “Noon”, and “Kaf”. (a) Recognition of “Ayn” using ResNet-18,(b) Recognition of “Ayn” using VGG-16,(c) Recognition of “Noon” using ResNet-18,(d) Recognition of “Noon” using VGG-16,(e) Recognition of” Kaf” using ResNet-18, and (f) Recognition of “Kaf” using VGG-16.

TABLE II. PERFORMANCE COMPARISON BETWEEN THE PROPOSED MODELS AND THE EXISTING WORK IN [34] AND [40].

Model	*Accuracy %*	*Precision %*	*Recall %*	*F1 %*
KNN & Skin-Profile based Approach [9]	14	13.2%	13.2	12.4
KNN & MLP based Approach [17]	41	40%	40.5	40.5
VGG_16 based Approach	93.2	93.3	94.3	93.7
ResNet-18 based Approach	93.4	93.6	93.5	93.5

To verify the performance of the skin-profile based approach [9], we tuned the number of neighbors from 1 to 200 with a step size of five. This proved that the number of neighbors is not the factor that affects recognition rate. To further illustrate the difference in performance, we show in Fig. 10 a sample image with a complicated background in which the signer has similar clothing and skin color, while our models were able to detect the hand and recognize the gesture with confidence of 0.63 and 0.58 using VGG-16 and ResNet respectively. The existing work [9] confused the clothing and the skin which lead to an incorrect classification.

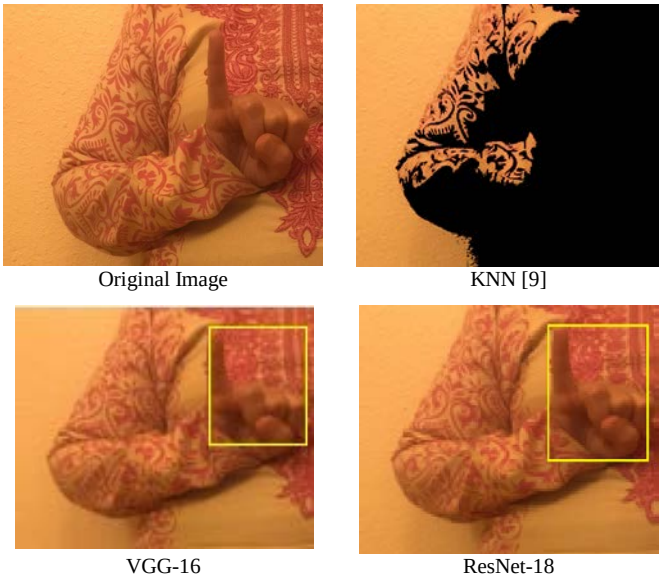

Fig. 10. Recognition of the Letter "Meem" using the Work in [9], R-CNN based on VGG-16 and R-CNN based on Resnet18.

The relatively poor performance of the existing work [9] can be attributed to the simple skin segmentation technique adopted by the authors. In fact, they adopted a YCbCr static skin segmentation which cannot handle different skin tones, lighting, and complex backgrounds. Moreover, the difference between the performance reported in [9] which attains 97%, and the one obtained using our dataset (14%) can be explained by the potential overfitting of their model when used with a very small dataset including 150 images only [9]. Furthermore, the second comparison was done with the work in [17] which outlined a substantial extension of the contributions in [9] that were affected by the considered skin segmentation technique. Specifically, a trained MLP was deployed to detect skin pixels in the images. Despite the ability to handle highly variant skin color and lighting, the system in [17] failed to distinguish the hand from a skin-colored background. In particular, the images that exhibit less complex background (i.e. non-skin color background) were correctly classified due to the three features extracted from the image. However, the majority of our dataset contains extremely complex background which yielded an accuracy of 41% for the work in [17]. Thus, one can claim that the experiments conducted in this research confirmed the ability of faster R-CNN to recognize efficiently the Arabic sign language. Moreover, they proved that the proposed system outperforms the relevant state of the art solutions in [9][17].

V. CONCLUSIONS

Arabic Sign Language is the primary form of communication within the Arab Deaf community. However, the sign language is not widely used and/or mastered outside this community which resulted in a real social barrier between Deaf and hearing people. In order to reduce this struggles for the Arab Deaf, researchers introduced ArSL recognition systems able to capture and recognize the hand gesture from images. Despite this effort, most of the reported works use datasets with uniform background in order to by-pass the image segmentation issue. Alternatively, ArSL recognition systems based on deep learning paradigms emerged to alleviate the concern of choosing the most relevant features. Taking into consideration the strengths and weaknesses of the state-of-the art contributions, we designed and implemented a novel ArSL recognition system that is able to localize and recognize the alphabet of the Arabic sign language using a Faster Region-based Convolutional Neural Network (R-CNN). Specifically, faster R-CNN was adapted to extract and map the image features, and learn the position of the hand in a given image. Moreover, the proposed system was assessed using a collection of 15,360 images, containing hand gestures with different backgrounds, captured using standard phone cameras. The association of the proposed architecture with ResNet and VGG-16 models achieved a recognition rate of 93% for the collected ArSL images dataset.

As future works, we propose to investigate the YOLO deep learning architecture [43] instead of Faster R-CNN for ArSL letter recognition. Unlike Faster R-CNN, YOLO can be adapted to conduct the classification and the bounding box regression simultaneously. It proved to achieve accurate and fast recognition when the objects of interest are not too small [43].

ACKNOWLEDGMENT

This work was supported by the Research Center of the college of Computer and information Sciences at King Saud University, Riyadh, KSA. The authors are grateful for this support.

REFERENCES

[1] Adam R. 2015. Standardization of Sign Languages. Sign Language Studies 15:432–445. DOI: 10.1353/sls.2015.0015.

[2] Ahmed MA, Zaidan BB, Zaidan AA, Salih MM, Lakulu and MM bin. A Review on Systems-Based Sensory Gloves for Sign Language Recognition State of the Art between 2007 and 2017. Sensors. DOI: 10.3390/s18072208.

[3] Neiva DH., Zanchettin C. 2018. Gesture recognition: A review focusing on sign language in a mobile context. Expert Systems with Applications 103:159–183. DOI: 10.1016/j.eswa.2018.01.051.

[4] Suharjito., Wiryana F., Kusuma GP., Zahra A. 2018. Feature Extraction Methods in Sign Language Recognition System: A Literature Review. 2018 Indonesian Association for Pattern Recognition International Conference (INAPR). DOI: 10.1109/inapr.2018.8626857.

[5] Abdel-Fattah MA. 2005. Arabic Sign Language: A Perspective. Journal of Deaf Studies and Deaf Education. DOI: 10.1093/deafed/eni007.

[6] Lecun Y., Bottou L., Bengio Y., Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86:2278–2324. DOI: 10.1109/5.726791.

[7] Egmont-Petersen M., Ridder DD., Handels H. 2002. Image processing with neural networks—a review. Pattern Recognition 35:2279–2301. DOI: 10.1016/s0031-3203(01)00178-9.

[8] Mohandes M., Liu J., Deriche M. 2014. A survey of image-based Arabic sign language recognition. 2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14). DOI: 10.1109/ssd.2014.6808906.

[9] Hemayed EE., Hassanien AS. 2010. Edge-based recognizer for Arabic sign language alphabet (ArS2V-Arabic sign to voice). 2010 International Computer Engineering Conference (ICENCO). DOI: 10.1109/icenco.2010.5720438.

[10] Prewitt J.M.S. 1971. Picture processing and psychopictorics. Icarus 15:563–564. DOI: 10.1016/0019-1035(71)90136-9.

[11] Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24:498–520. DOI: 10.1037/h0070888.

[12] Serra J. 1993. Image analysis and mathematical morphology. London: Academic.

[13] Altman NS. 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician 46:175. DOI: 10.2307/2685209.

[14] Hamed A., Belal NA., Mahar KM. 2016. Arabic Sign Language Alphabet Recognition Based on HOG-PCA Using Microsoft Kinect in Complex Backgrounds. 2016 IEEE 6th International Conference on Advanced Computing (IACC). DOI: 10.1109/iacc.2016.90.

[15] Cortes C., Vapnik V. 1995. Machine Learning 20:273–297. DOI: 10.1023/a:1022627411411.

[16] Velho L., Frery AC., Gomes J., Gomes J. 2009. Image processing for computer graphics and vision. New York, NY: Springer.

[17] Dahmani D., Larabi S. 2014. User-independent system for sign language finger spelling recognition. Journal of Visual Communication and Image Representation 25:1240–1250. DOI: 10.1016/j.jvcir.2013.12.019.

[18] Rosenblatt F. 1961. Principles Of Neurodynamics. Perceptrons And The Theory Of Brain Mechanisms. DOI: 10.21236/ad0256582.

[19] Hu M-K. 1962. Visual pattern recognition by moment invariants. IEEE Transactions on Information Theory 8:179–187. DOI: 10.1109/tit.1962.1057692.

[20] El-Bendary N., Zawbaa HM., Daoud MS., Hassanien AE., Nakamatsu K. 2010. ArSLAT: Arabic Sign Language Alphabets Translator. 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM). DOI: 10.1109/cisim.2010.5643519.

[21] Lowe DG. 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60:91–110.

[22] Tharwat A., Gaber T., Hassanien AE., Shahin MK., Refaat B. 2015. SIFT-Based Arabic Sign Language Recognition System. Advances in Intelligent Systems and Computing Afro-European Conference for Industrial Advancement:359–370. DOI: 10.1007/978-3-319-13572-4_30.

[23] Fisher RA. 1936. The Use Of Multiple Measurements In Taxonomic Problems. Annals of Eugenics 7:179–188. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

[24] Al-Jarrah O., Halawani A. 2001. Recognition of gestures in Arabic sign language using neuro-fuzzy systems. Artificial Intelligence 133:117–138. DOI: 10.1016/s0004-3702(01)00141-2.

[25] Jang J-S. 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics 23:665–685. DOI: 10.1109/21.256541.

[26] Sadeddine K., Djeradi R., Chelali FZ., Djeradi A. 2018. Recognition of Static Hand Gesture. 2018 6th International Conference on Multimedia Computing and Systems (ICMCS). DOI: 10.1109/icmcs.2018.8525908.

[27] Canny J. 1987. A Computational Approach to Edge Detection. Readings in Computer Vision:184–203. DOI: 10.1016/b978-0-08-051581-6.50024-6.

[28] He D-C., Wang L. 1990. Texture Unit, Texture Spectrum And Texture Analysis. 12th Canadian Symposium on Remote Sensing Geoscience and Remote Sensing Symposium. DOI: 10.1109/igarss.1989.575836.

[29] Zernike VF. 1934. Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode. Physica 1:689–704. DOI: 10.1016/s0031-8914(34)80259-5.

[30] Specht DF. 1990. Probabilistic neural networks. Neural Networks 3:109–118. DOI: 10.1016/0893-6080(90)90049-q.

[31] Alzohairi R., Alghonaim R., Alshehri W., Aloqeely S. 2018. Image based Arabic Sign Language Recognition System. International Journal of Advanced Computer Science and Applications 9. DOI: 10.14569/ijacsa.2018.090327.

[32] Hayani S., Benaddy M., Meslouhi OE., Kardouchi M. 2019. Arab Sign language Recognition with Convolutional Neural Networks. 2019 International Conference of Computer Science and Renewable Energies (ICCSRE). DOI: 10.1109/iccsre.2019.8807586.

[33] Jain LC., Medsker L. 2000. Recurrent neural networks: design and applications. Boca Raton, FL: CRC Press.

[34] Maraqa M., Abu-Zaiter R. 2008. Recognition of Arabic Sign Language (ArSL) using recurrent neural networks. 2008 First International Conference on the Applications of Digital Information and Web Technologies (ICADIWT). DOI: 10.1109/icadiwt.2008.4664396.

[35] Zell A. 2003. Simulation neuronaler Netze. München: Oldenbourg.

[36] Elman JL. 2020. Finding structure in time. Connectionist psychology: A text with readings:289–312. DOI: 10.4324/9781315784779-11.

[37] Jordan MI. 1986. Serial order: a parallel distributed processing approach. La Jolla, CA: Institute for Cognitive Science, University of California, San Diego.

[38] Dunn JC. 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics 3:32–57. DOI: 10.1080/01969727308546046.

[39] Aly S., Osman B., Aly W., Saber M. 2016. Arabic sign language fingerspelling recognition from depth and intensity images. 2016 12th International Computer Engineering Conference (ICENCO). DOI: 10.1109/icenco.2016.7856452.

[40] Ren S, He K, Girshick R, Sun J. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/tpami.2016.2577031.

[41] Simonyan K, Zisserman A. 2015. Very Deep Convolutional Networks for Large-Scale. ICLR.

[42] He K., Zhang X., Ren S., Sun J. 2016. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). DOI: 10.1109/cvpr.2016.90.

[43] Redmon J., Divvala S., Girshick R., Farhadi A. 2016. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). DOI: 10.1109/cvpr.2016.91.

[44] Ethnologue. 2019.Languages of the World. Available at http://www.ethnologue.com/15/show_family/90008/ (accessed April 1, 2019).