Fayoum University Faculty of Engineering, 2024, Vol: 7(1), 40-48
https://fuje.journals.ekb.eg/
ISSN Online: 2537-0634
ISSN Print: 2537-0626

A Novel Image-Based Arabic Hand Gestures Recognition Approach Using YOLOv7 and ArSL21L

Fatma M. A. Mazen[1,*], Mai Ezz-Eldin [1,2]

[1]Electronics and Communication Engineering Department, Faculty of Engineering, Fayoum University, Fayoum, Egypt, 63514

[2]Electrical Engineering Department, Future High Institute for Engineering, Fayoum, Egypt

*Corresponding author: Fatma M. A. Mazen (fma04@fayoum.edu.eg).

How to cite this paper: Mazen, F.M.A, and Ezz-Eldin, M (2024). A Novel Image-Based Arabic Hand Gestures Recognition Approach Using YOLOv7 and ArSL21L., *Fayoum University Journal of Engineering, Vol: 7(1), 40-48*
https://dx.doi.org/10.21608/FUJE.2023.216182.1050

Abstract

Recognizing and documenting Arabic sign language has recently received a lot of attention because of its ability to enhance communication between deaf persons and normal people. The development of automatic sign language recognition (SLR) systems to allow communication with deaf persons is the primary goal of SLR. Until recently, Arabic SLR (ArSLR) received little attention. Building an automatic Arabic hand gesture recognition system is a challenging task. This work presents a novel image-based ArSL recognition approach where You Only Look Once v7 (YOLOv7) is used to build an accurate ArSL alphabet detector and classifier utilizing ArSL21L: Arabic Sign Language Letter Dataset. The proposed YOLOv7 medium model has achieved the highest mAP0.5 and mAP0.5:0.95 scores of 0.9909 and 0.8306, respectively. It has outperformed not only YOLOv5m but also YOLOv5l in terms of mAP0.5 and mAP0.5:0.95 scores. Furthermore, regarding mAP0.5 and mAP0.5:0.95 scores, the YOLOv7-tiny model has not only surpassed YOLOv5s but additionally YOLOv5m. YOLOv5s, on the other hand, has the lowest mAP0.5 and mAP0.5:0.95 scores of 0.9408 and 0.7661, respectively.

Keywords

Keywords: Arabic Sign Language Letter Dataset (ArSL21L); Convolutional Neural Network; YOLOv7; YOLOv5; Arabic Sign Language Recognition (ArSLR)

1. Introduction

Sign language is the principal way of interaction and communication between deaf people with others. In addition to facial expressions, sign language utilizes manual illustration and the movement and position of body parts like hands, arms, and wrists to reflect the signer's ideas and feelings [1]. Sensor-based and image-based techniques are the two basic methodologies for SLR [2]. While being the image-based approach user friendly since it requires only cameras for capturing signs (gestures), the sensor-based approach requires high computational cost since users have to wear devices like motion trackers or data gloves. Furthermore, it might be quite sensitive to variations in lighting and background conditions. Methods for recognizing the ArSL are usually categorized into three levels: identifying the hand movements of the Arabic alphabet at the word level, which includes recognizing single motions, and sentence level, which provides for recognizing continuous hand movements [3]. This research will focus on recognizing Arabic sign language alphabets. As a result, we will spotlight more on image-based systems, particularly those dealing with identifying alphabets in Arabic sign language.

2. Related Work

This section discusses the latest classification and detection studies for Arabic sign language. The authors of [4] have tried several pre-trained models. Considering the size of the ArSL2018 dataset applied to all of the proposed models, all but EfficientNetB4 perform poorly because of their lightweight architecture. EfficientNetB4 is a massive architecture with more complexities in comparison. Nevertheless, the EfficientNetB4 model achieved 98% training and 95% testing accuracy.

With paying more attention to image-based systems, specifically ones relating to recognizing alphabets in Arabic sign language, we must spot on Zabulisy et al. [5]. He proposed an efficient and robust image-based hand gesture recognition system that allows interaction with autonomous robots that guide tourists around museums and exhibition sites.

On the other hand, Tubaiz et al. [6] developed a system for sequential Arabic Sign Language Classification. To simulate dynamic continuous signs, they used two DG5-VHand data gloves and a camera. The accumulated data was refined with re-sampling and z-score normalization before being classified with a modified KNN. They achieved a 98.90% accuracy in sentence recognition. Utilizing the same dataset, their system not only outperformed a prior vision-based system by 23.9%, but it also overcomes the limitations of vision-based systems.

In another study [7], Sarhan et al. used Kinect to develop an Arabic Sign Language identification system. The primary goal was to develop a robust system to help communication between a hard-of-hearing or deaf patient and the doctor's medical hospitals. Due to the lack of an existing dataset for ArSL words, a new one was created utilizing Kinect. They collected 215 dynamic sample data of hand movements, hand shapes, and articulation points for 16 Arabic words, which were subsequently obtained using depth and skeletal data. The Hidden Markov Model (HMM) was used for classification with ten-fold cross-validation, achieving an accuracy of 80.47%.

Hassan et al. [8] created a continuous sensor-based Arabic Sign Language Recognition system. They used a Polhemus G4 tracker and DG5-VHand data gloves to collect two datasets containing 40 Arabic sentences totaling 80 words. A different signer was responsible for collecting each dataset. A sliding window was used to extract the features, which were then classified using the Modified K Nearest Neighbors (MKNN) and HMM approaches. MKNN outperformed HMM in sentence classification and performed exceptionally well, with a 97% sentence recognition rate, while HMM outperformed in word classification.

The authors in [9] and [10] utilized a Support Vector Machine (SVM) with a Histogram of Oriented Gradients

(HOG) to automatically recognize ArSL alphabets. The proposed system achieved a recognition accuracy equal to 63.5% and 99.2% of Arabic Alphabet gestures, respectively.

Ghazanfar Latif et al. [11] provided the first open dataset for recognizing Arabic Alphabets Sign Language (ArSL2018). The authors collected the ArSL2018 dataset at Prince Mohammad Bin Fahd University and Al Khobar, Saudi Arabia, from 40 different age group participants. The dataset includes 54049 grayscale photos of 32 ArSL alphabets taken for one hand with white background at a pixel density of 64x64. Moreover, Althagafi et al. [12] utilized the ArSL2018 dataset to design an ASLR system to recognize the Arabic language's alphabet signs. They trained and tested a custom Convolutional Neural Network (CNN) model using 80% and 20% of the dataset, respectively. They reached a recognition rate of 94.46%.

The following sections are in order: Section 2 describes the dataset. Section 3 describes the proposed methodology and the architecture of the suggested model. Section 4 discusses the outcomes and evaluation. Finally, Section 5 provides a conclusion.

3. The Dataset

It is challenging for deaf people to make social connections with people as this requires interactive systems that can recognize sign language. There are numerous datasets and methods for English sign language (SL); regrettably, Arabic sign language (SL) is severely limited. Arabic Sign Language Letter Dataset (ArSL21L) [13] includes annotated Arabic Sign Language Letters. As shown in Table 1, the ArSL21L database comprises 14202 images of 32 letter signs with diverse backgrounds gathered from 50 volunteers, 23 females, and 27 males.

The volunteers ranged in age from 14 to 69, with a standard deviation of 13.99. Samples of these volunteers are illustrated in Figure 1. There are nine images for each sign in the ArSL21L dataset, taken with volunteers' smartphones, with three shots from different angles for each distance. As a result, each signer received a total of 288 images; however, several poorly qualified photos were excluded during the dataset's formation. Signers could use either hand to make the signs. All of the images were resized to 416x416 pixels. The total number of images in the ArSL21 dataset is 14202 images of 32 different signs with various of backgrounds. The dataset was annotated with bounding boxes in PASCAL Visual Object Classes (PASCAL VOC) [14] format using the Labeling program [15].

Table1.: Signs and images recordings details of ArSL21L dataset

Dataset Parameters	Description
No. Classes	32
No. of signers	50
Total images	14202
Training images	9955 (35 signers)
Testing images	4247 (15 signers)
Mean samples per class	443.8
Resolution	416x416

Figure 1: Dataset samples

4. Methods

In this work, we utilize the latest YOLO algorithm, YOLOv7. The YOLOv7 algorithm is causing quite a stir in the machine learning and computer vision communities. It significantly outperformed all prior YOLO versions and object detection models regarding accuracy and speed. Rather than employing pre-trained ImageNet backbones, YOLOv7 models are trained using the COCO dataset [16]. The authors of YOLOv7

are the same as the authors of Scaled YOLOv4. The significant difference between YOLOv7 over the scaled YOLOv4 is that YOLOv7 utilizes an Extended Efficient Layer Aggregation Network (E-ELAN). The E-ELAN architecture allows the framework to learn more effectively. The YOLOv7 also proposed a Compound Model Scaling technique. The authors of YOLOv7 have introduced some Bag of Freebies (BoF) techniques to help improve the model's performance without increasing training costs, like Planned re-parameterized convolution, Coarse for Auxiliary, and Fine for Lead loss. Re-parameterization is used after training to boost the model's inference results. Model level and Module level ensemble re-parametrization are the two categories used to finalize models. The Lead Head in YOLOv7 is the head in charge of the final output. The Auxiliary Head is the head that assists training in the middle layers and and boosts the model learning. YOLOv7 models include the YOLOv7 version optimized for standard GPU computing, YOLOv7tiny optimized for edge GPU, and YOLOv7-W6 cloud GPU computing. Contrary to all versions of the YOLOv7 model, which use SiLU as the activation function, the YOLOv7-tiny utilizes leaky ReLU. In this research, due to hardware limitations, we only utilized the YOLOv7 and YOLOv7-tiny models. The network architecture of YOLOv7 is depicted in Fig.2. According to the structural diagram, the YOLOv7 network comprises three components, namely the input network, backbone network, and head network. The input network preprocesses the image by resizing it to 640 x 640 x 3 before feeding it into the backbone network. The Cross mini-Batch Synchronization (CBS) composite module, efficient layer aggregation networks (ELAN) module, and MP module are utilized to iteratively reduce the length and width of the feature map by a factor of 1/2. Additionally, the number of output channels is doubled compared to the number of input channels.

To commence the training process, a set of hyper-parameters needed to be established. The training was conducted using a Kaggle GPU P100, as it offered sufficient resources for the task. The system specifications, along with the defined hyper-parameters employed to guide the training process, are presented in Table 2.

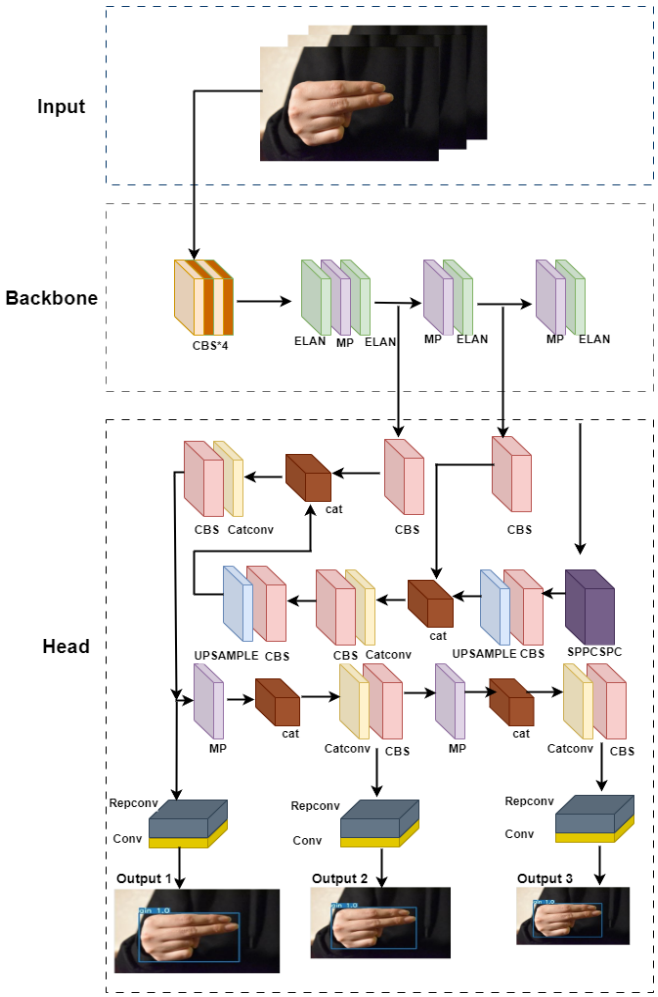

Fig.2: Architecture of YOLOv7 model

Table2: Hyper-parameter setting for the YOLOv7 model

Hyper-parameter	Value
Batch Size	16
Epochs	100
Optimizer	Adam
Image Size	640×640
lr0	0.01
lrf	0.1
momentum	0.937
weight_decay	0.0005
warmup_epochs	3.0
warmup_momentum	0.8
warmup_bias_lr	0.1
GPU Memory	16 GB
GPU	Kaggle GPU P100

5. Results and Discussion

5.1. Evaluation Metrics

The metrics employed to assess the performance of the suggested Arabic sign language detection models are recall, precision, and mean Average Precision (mAP0.5 and mAP 0.5-0.95). Precision is an index of exactness that defines the percentage of false positives in the dataset. Conversely, recall measures a match's goodness and a model's effectiveness in identifying positive labels [17]. The mAP computes a score by comparing the predicted box to the reference bounding box. A higher score is an indicator of a better detection model. The equations [1], [2], and [3] denote Precision (*P*), Recall (*R*), and mean Average Precision (*mAP*) respectively:

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_j \quad (3)$$

where: $AP = \int_0^1 P(R)dR$, TP= True Positive, FP= False Positive, FN= False Negative, n= the number of classes.

5.2. Performance Evaluation

The performance evaluation metrics of YOLOv7-tiny is summarizes in Table 3. YOLOv7 medium models, and different YOLOv5 models trained on the ArSL21L dataset, like Recall, Precision, and mAP values. YOLOv7 medium model has the highest mAP0.5 and mAP0.5:0.95 scores of 0.9909 and 0.8306, respectively, of the models. It has outperformed not only YOLOv5m but also YOLOv5l in terms of mAP0.5 and mAP0.5:0.95 scores. Furthermore, in terms of mAP0.5 and mAP0.5:0.95 scores, the YOLOv7-tiny model has not only surpassed YOLOv5s but also YOLOv5m. YOLOv5s, on the other hand, has the lowest mAP0.5 and mAP0.5:0.95 scores of 0.9408 and 0.7661, respectively.

Table 3: Evaluation of YOLOv7 tiny and medium models on the validation set of ArSL21 dataset

Model	Precision	Recall	mAP 0.5	mAP 0.5:0.95
YOLOv5s	0.953	0.9408	0.9784	0.7661
YOLOv5m	0.968	0.9468	0.9842	0.7768
YOLOv7-tiny[ours]	**0.964**	**0.964**	**0.988**	**0.811**
YOLOv5x	0.9758	0.9743	0.9896	0.8224
YOLOv5l	0.9787	0.9766	0.9909	0.8306
YOLOv7-medium[ours]	**0.982**	**0.983**	**0.992**	**0.836**

For various threshold values, the Precision-Recall curve (PR curve) represents the tricky balance between recall and precision [5]. AP is the area under the PR curve in Equation (9), and mAP is the average of APs from different classes. The number of test sample classes is denoted by N. N = 32 because the dataset contains 32 Arabic Alphabet Sign Language categories. The greater the area under the curve, the higher the precision and recall values. High precision is associated with a low false positive rate, whereas high recall is associated with a low false negative rate. The Precision-Recall curves for the YOLOv7 tiny and medium models are shown in Fig.3, respectively.

For training, the training and validation sets were fed into the network. Fig.4, depicts the evaluation metrics and the loss function value curves of the training and validation sets after 100 batches of training. There are three types of losses in the loss curves: detection frame (box) loss, detection object (objectness) loss, and classification loss. The box loss indicates whether an algorithm can precisely locate an object's center point and whether the detection target is surrounded by the predicted bounding box. The more accurate the prediction bounding box, the smaller the loss function value. The object ness loss function is a probability measure of the detection

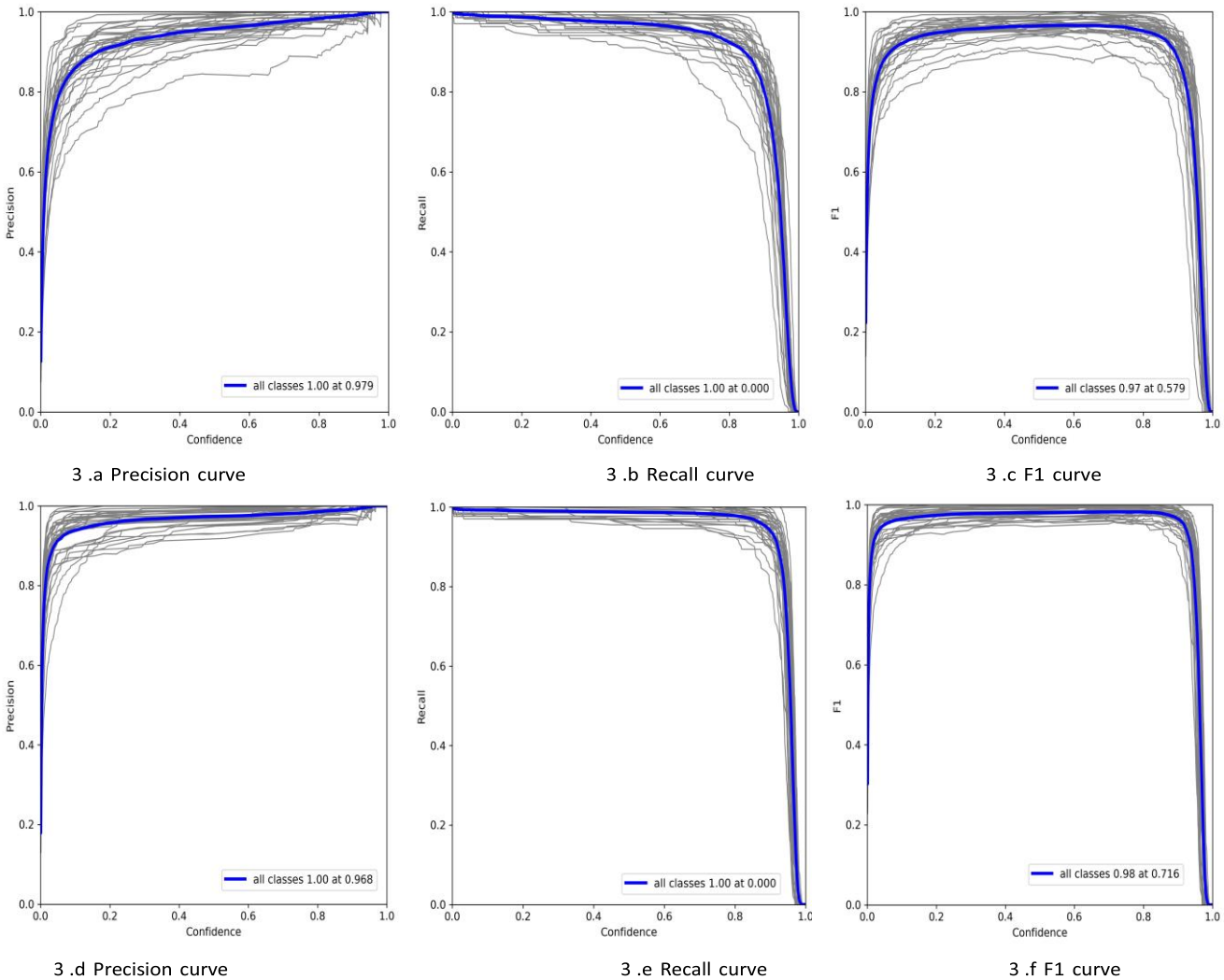

Fig.3: illustrates the Precision, Recall and F1 curves. Figures 3.a, 3.b, and 3.c for Tiny model where 3.d, 3.e, and 3.f for Medium model respectively

target's presence in the region of interest. The higher the accuracy, the smaller the value of the loss function. The classification loss denotes the model's capability to correctly identify a given object category. The lower the loss value, the better the classification. Validation metrics reveal that the model's performance gradually improved throughout the training process, which indicates that the model converged quickly and produced good results. Following training, the model was tested using the ArSL21 dataset's test set. Fig.5 depicts some examples of Arabic alphabet sign language detection. The left column represents a batch of test images of the ArSL21 dataset with ground truth bounding boxes labeled "Ground truth batch." The same batch of images is shown in the right column but with the bounding boxes predicted by the proposed YOLOv7 medium model.

6. Conclusions

This paper presents a comprehensive outline of the design of an Arabic Sign Language Detection system utilizing YOLOv7, discussing network architecture, parameter settings, and the Arabic Sign Language Letter Dataset (ArSL21L). The study involved experiments with YOLOv7

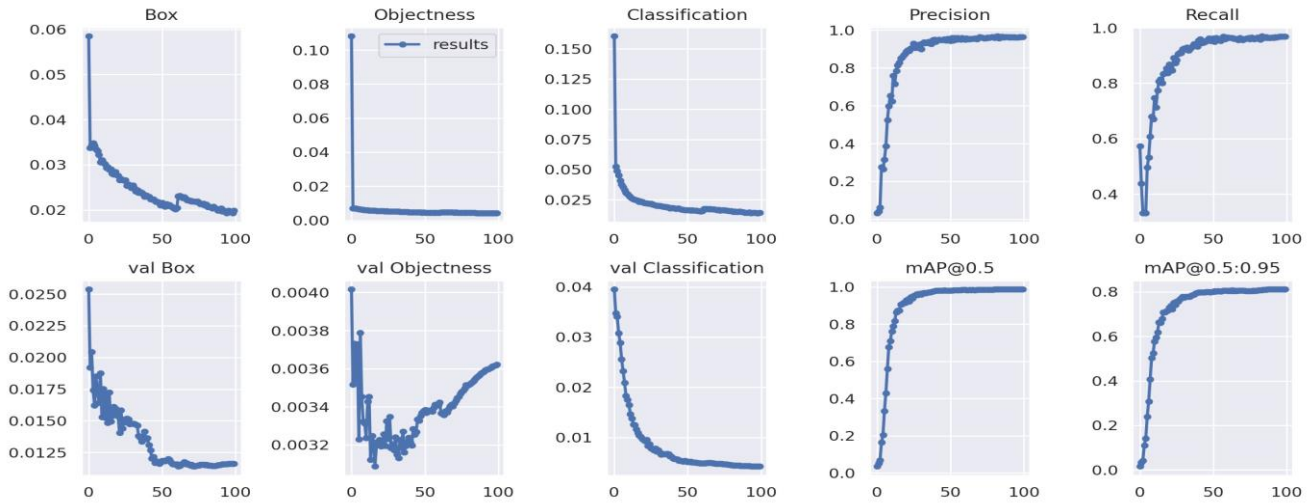

Fig.4.a Losses and Evaluation metrics for the proposed Tiny model

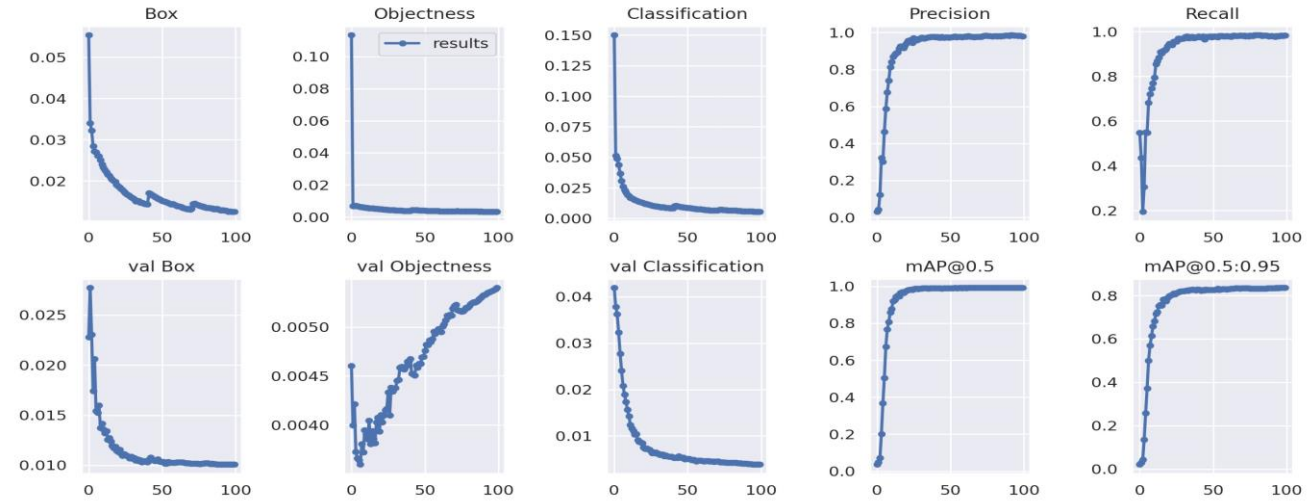

Fig.4.b Losses and Evaluation metrics for the proposed Medium model

Fig.4: 4 .a and 4.b illustrate the Evaluation metrics and Losses for proposed Tiny and Medium models, respectively

models (T, m) and comparisons to YOLOv5 models (s, m, l, x). The results indicate that the YOLOv7-based models demonstrated superior performance compared to the YOLOv5-based models in terms of Precision, Recall, mAP0.5, and mAP0.5:0.95 scores. Efforts toward future research will strive to extend this proposed alphabet detector for Arabic Sign Language to facilitate word and sentence recognition

7. References

[1] B. Hisham, A. Hamouda, Arabic sign language recognition using ada-boosting based on a leap motion controller, International Journal of Information Technology 13 (2021) 1221–1234.

[2] M. Hassan, K. Assaleh, T. Shanableh, Multiple proposals for continuous arabic sign language recognition, Sensing and Imaging 20 (2019) 1–23.

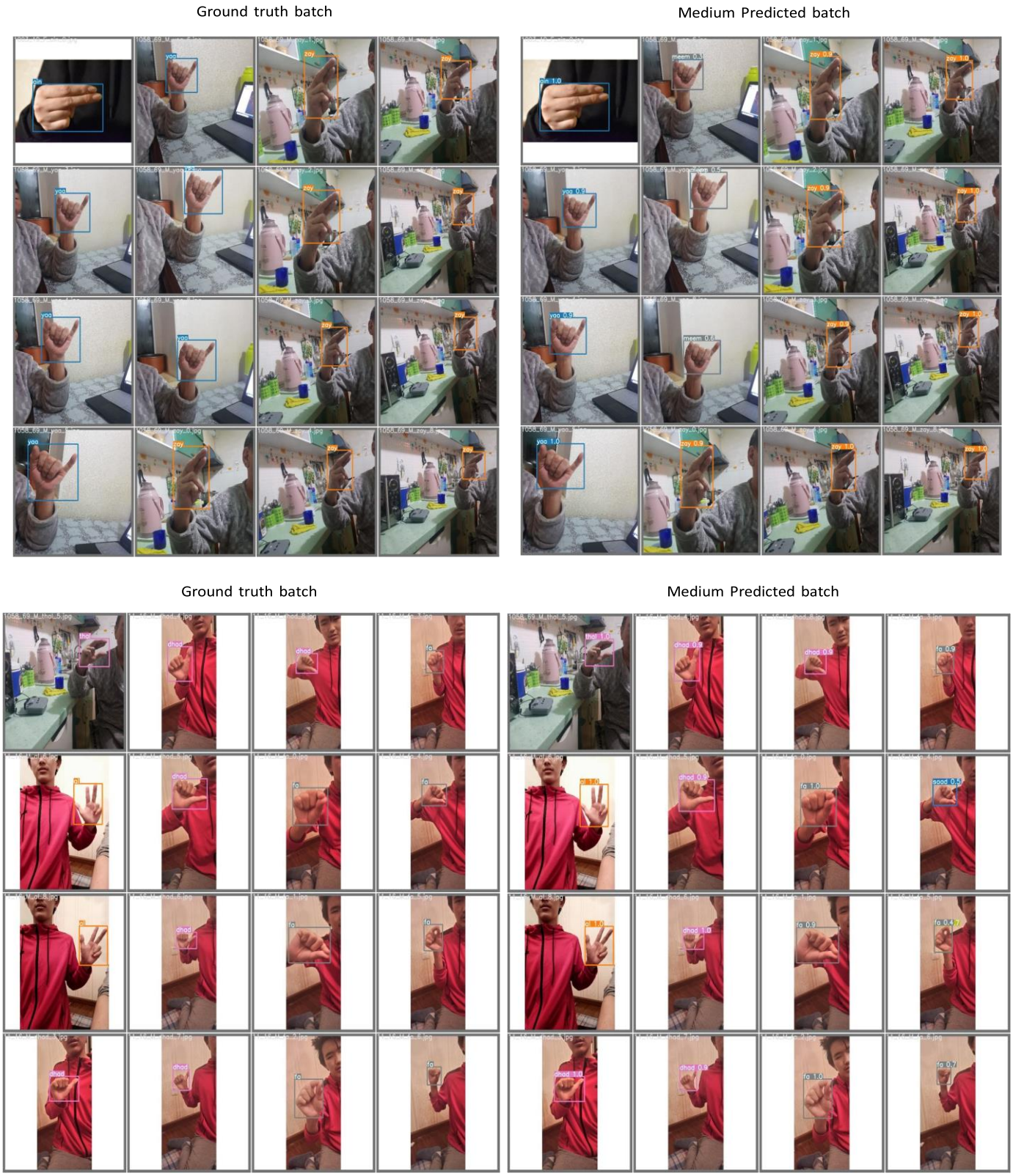

Fig.5: Samples of ground truth bounding boxes vs predicted bounding boxes for proposed medium model

[3] G. Tharwat, A. M. Ahmed, B. Bouallegue, Arabic sign language recognition system for alphabets using machine learning techniques, Journal of Electrical and Computer Engineering 2021 (2021) 1–17.

[4] M. Zakariah, Y. A. Alotaibi, D. Koundal, Y. Guo, M. Mamun Elahi, Sign language recognition for arabic alphabets using transfer learning technique, Computational Intelligence and Neuroscience 2022 (2022).

[5] X. Zabulis, H. Baltzakis, A. A. Argyros, Vision-based hand gesture recognition for human-computer interaction., The universal access handbook 34 (2009) 30.

[6] N. Tubaiz, T. Shanableh, K. Assaleh, Glove-based continuous arabic sign language recognition in user-dependent mode, IEEE Transactions on Human-Machine Systems 45 (4) (2015) 526–533.

[7] Sarhan, N. A., El-Sonbaty, Y., & Youssef, S. M. (2015, October). HMM-based Arabic sign language recognition using kinect. In 2015 tenth international conference on digital information management (ICDIM) (pp. 169-174). IEEE.

[8] M. Hassan, K. Assaleh, T. Shanableh, User-dependent sign language recognition using motion detection, in: 2016 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, 2016, pp. 852–856.

[9] R. Alzohairi, R. Alghonaim, W. Alshehri, S. Aloqeely, Image based arabic sign language recognition system, International Journal of Advanced Computer Science and Applications 9 (3) (2018).

[10] A. Hamed, N. A. Belal, K. M. Mahar, Arabic sign language alphabet recognition based on hog-pca using microsoft kinect in complex backgrounds, in: 2016 IEEE 6th international conference on advanced computing (IACC), IEEE, 2016, pp. 451–458.

[11] G. Latif, N. Mohammad, J. Alghazo, R. AlKhalaf, R. AlKhalaf, Arasl: Arabic alphabets sign language dataset, Data in brief 23 (2019) 103777.

[12] E. Aldhahri, R. Aljuhani, A. Alfaidi, B. Alshehri, H. Alwadei, N. Aljojo, A. Alshutayri, A. Almazroi, Arabic sign language recognition using convolutional neural network and mobilenet, Arabian Journal for Science and Engineering 48 (2) (2023) 2147–2154.

[13] G. Batnasan, M. Gochoo, M.-E. Otgonbold, F. Alnajjar, T. K. Shih, Arsl21l: Arabic sign language letter dataset benchmarking and an educational avatar for metaverse applications, in: 2022 IEEE Global Engineering Education Conference (EDUCON), IEEE, 2022, pp. 1814–1821.

[14] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2010) 303–338.

[15] D. Tzutalin, Labelimg, GitHub repository 6 (2015).

[16] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.

[17] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information processing & management 45 (4) (2009) 427–437.