University of Mosul
College of Engineering

Arabic Sign Language Detection and Recognition Using Deep Learning Based-Heterogeneous Computing

Mohammad Haqqi Ismail

Ph.D. Thesis
In Computer Engineering

Supervised By

Dr. Shefa Abdulrahman Dawwd
Professor

Dr. Fakhrulddin Hamid Ali
Assistant Professor

2022 A.D. 1444 A.H.

ABSTRACT

Automatic detection of sign language from hand gesture video streams is critical. Deaf and mute people may benefit from accurate detection and recognition of sign language. When processing a series of video frames captured by the camera, it is necessary to find out if the person is signing or not. The present study aims to develop a model for detecting signers in a video stream for Arabic sign language in real time and to classify the signs among static, dynamic, and non-sign. The first step in preprocessing the data is extracting keypoints of poses in video frames using the MediaPipe library. By MediaPipe, the pose detection algorithm first detects the location of the human in the video using a CNN. And then, a secondary CNN was used to predict the human keypoints. The best architecture design model used for Sign Detection of Arabic sign language classifies the signs among static, dynamic, and non-sign with training accuracy of 100% and test accuracy of 99.05%. The model used a limited number of keypoints selected from the body pose and a special normalization approach to extract features. The model included two bidirectional layers using Gated Recurrent Unit (GRU) for both forward and backward layers and ended with the fully connected layer for classification using SoftMax.

For Arabic static sign language recognition in real time, a pre-trained models were used and trained on prepared Arabic Sign Language data. These models were used after some modification. Also, an attempt has been made to adopt two models from the previously trained models, where they are trained in Parallel Deep Feature Extractions. Then they are combined and prepared for the classification stage. The results demonstrate the comparison between the performance of the single model and the multi-model. It appears that most multi-models are better in feature extraction and classification than single models. The results also show that depending on the total number of Incorrect Recognize sign image in the training, validation, and testing dataset, the best CNN model in feature extraction and classification of Arabic sign language is the

DenseNet121 for a single model and DenseNet121 & VGG16 for multi-model.

For Arabic dynamic sign language recognition in real time, four deep neural network models were proposed using 2D and 3D CNN to cover all feature extraction methods and then passing these features to the RNN for sequence classification. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) which are two types of Recurrent Neural Network (RNN) are used. The study also included evaluation fusion techniques for various kinds of multiple models. The experiment results show the best multi-model for the dynamic dataset of the Arabic sign language recognition achieved 100% accuracy.

A set of models used in the detection and recognition of sign language gestures are proposed to increase the speed of inference by the concept of optimization using the TensorRT inference accelerator tool. This is performed using different devices: Laptop, Jetson Xavier Nx, and Cloud Computing. When TF-TRT (TensorFlow-TensorRT) Integration optimization is applied, the speed of inference increases about 3x-11x for static sign language recognition models and about 1x-3x for dynamic sign recognition models. On the other hand, using optimization by applying TensorRT (TensorRT C++ API), the speed of inference increases about 14x-110x for static sign language recognition models and about 2x- 10x for dynamic sign recognition models.

جامعة الموصل
كلية الهندسة

كشف وتمييز لغة الاشارة العربية باستخدام التعلم العميق للحوسبة غير المتجانسة

محمد حقي اسماعيل

اطروحة دكتوراه فلسفة
في هندسة الحاسوب

بإشراف

الأستاذ الدكتور
شفاء عبد الرحمن داؤد

الأستاذ المساعد الدكتور
فخر الدين حامد علي

محرم 1444هـ

آب 2022 م

المستخلص

يعد الاكتشاف التلقائي للغة الإشارة من تدفقات فيديو إيماءات اليد أمرًا بالغ الأهمية. قد يستفيد الأشخاص الصم والبكم من الاكتشاف الدقيق والتعرف على لغة الإشارة. عند معالجة سلسلة من إطارات الفيديو التي تم التقاطها بواسطة الكاميرا ، من الضروري معرفة ما إذا كان الشخص يقوم بتادية اشارة أم لا. تهدف الدراسة المقدمة إلى تطوير نموذج لكشف الإشارات في تدفق فيديو للغة الإشارة العربية في الزمن الحقيقي وتصنيف الإشارات بين ساكن وديناميةولايوجد إشارة. تتمثل الخطوة الأولى في المعالجة المسبقة للبيانات في استخراج النقاط الأساسية للموقع في إطارات الفيديو باستخدام مكتبة MediaPipe. بواسطة MediaPipe، خوارزمية كشف الموقع أولاً موقع الإنسان في الاطار الفيديوي باستخدام CNNوبعد ذلك ، ثانيا تم استخدام شبكة CNN ثانوية للتنبؤ بالنقاط الأساسية للشخص. أفضل نموذج تصميم معماري مستخدم في الكشف عن الإشارات بلغة الإشارة العربية يصنف الإشارات بين ساكن وديناميولايوجد إشارة مع دقة تدريب 100٪ ودقة اختبار 99٪. استخدم النموذج عددًا محدودًا من النقاط الأساسية المختارة من موقع الجسم واسلوب خاص للتطبيع لاستخراج الميزات. يتضمن النموذج طبقتين ثنائية الاتجاه باستخدام GRU لكل من الطبقات الأمامية والخلفية وانتهى بالطبقة المتصلة بكثافة للتصنيف باستخدام SoftMax.

من أجل التعرف على لغة الإشارة العربية الساكنة في الزمن الحقيقي، تم استخدام النماذج المدربة مسبقًا واعادة تدريبها على بيانات لغة الإشارة العربية المعدة. تم استخدام هذه النماذج بعد بعض التعديلات. كذلك، تم إجراء محاولة لاعتماد نموذجين من النماذج التي تم تدريبها سابقا ، حيث يتم تدريبهم على استخراج الميزات العميقة المتوازية. ثم يتم دمجها وتحضيرها لمرحلة التصنيف. بينت النتائج المقارنة بين أداء النموذج الفردي والنموذج المتعدد. ويبدو أن معظم النماذج المتعددة أفضل في استخراج الميزات والتصنيف مما عليه من النماذج الفردية. وكذلك أظهر أيضًا أنه بناءً على العدد الكلي لصورة الاشارة المميزة بشكل خطأ في مجموعة بيانات التدريب والتحقق والاختبار، أفضل نموذج لشبكة CNN في

استخراج الميزات وتصنيف لغة الإشارة العربية هو DenseNet121 باستخدام النموذج المفرد و DenseNet121 & VGG16 باستخدام النموذج المتعدد.

من أجل التعرف على لغة الإشارة العربية الدينامية في الزمن الحقيقي، تم اقتراح أربعة نماذج للشبكات العصبية العميقة باستخدام 2D و CNN 3D لتغطية جميع طرق استخراج الميزات ثم تمرير هذه الميزات إلى RNN لتصنيف المتسلسل . LSTM و GRU نوعين من استخدامRNN. كما تضمنت الدراسة تقييم تقنيات الاندماج لأنواع مختلفة من النماذج المتعددة. أظهرت نتائج التجربة أن النموذج المتعدد الأمثل لمجموعة البيانات الدينامية الخاصة بالتعرف على لغة الإشارة العربية قد حقق دقة 100٪.

تم اقتراح مجموعة من النماذج المستخدمة في الكشف عن إيماءات لغة الإشارة والتعرف عليها لزيادة سرعة الاستدلال من خلال مفهوم التحسين باستخدام أداة تسريع الاستدلال TensorRT. يتم ذلك باستخدام أجهزة مختلفة: الحاسوب الشخصي، و Jetson Xavier Nx، والحوسبة السحابية. عند تطبيق تحسين TF-TRT، تزداد سرعة الاستدلال بحوالي 11x-3x لنماذج التعرف على لغة الإشارة الساكنة وحوالي 3x-1x لنماذج التعرف الدينامية على الإشارة. أثناء استخدام تحسين TensorRT المطبق TensorRT) (API ++ C، تزداد سرعة الاستدلال بحوالي 110x-14x لنماذج التعرف على لغة الإشارة الساكنة وحوالي 10x- 2x لنماذج التعرف الدينامية على الإشارة.