Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science

Lip Shape Extraction and Movement for Arabic Speech Recognition Using Deep Learning Techniques

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Doctor of Philosophy in
Computer Science**

By

Nagham Salim Mohammed Allella

Supervised by

Prof. Dr. Khalil Ibrahim Al Saif

2023 A.D. **1444 A.H.**

Abstract

Lip reading is used to understand or interpret speech without hearing, its a technique especially mastered by people with hearing difficulties. The ability to lip read enables a person with a hearing impairment to communicate with others and to engage in social activities, which otherwise would be difficult.

In this research proposed deep learning and computer vision techniques for Arabic visual speech motion lips, the dataset was collected by recording silent video using webcam, because didn't find previous public Arabic dataset, (9) words recorded for (10) different talked participant

An Arabic Visual Speech Recognition system was designed and built consists of several main and sub-stage for data processing and obtain the highest recognition accuracy, after collected data Histogram Oriented Gradients (HOG) method to pinpoint the location of the face ROI extracted using mouth facial landmarks algorithms.

The first proposed system suggested for tracking lips movement figure out the lips shape and for this using Gaussian Filter, Hue Saturation Value (HSV), Morphological Operation and finally applying Canny Filter, to extracted and isolated lip border as visual speech feature, then for tracking lips movement fitted lips contour into polynomial function with eight order and its parameters determine the shape of the lip movement curve, which is used to calculate the derivative of lip movement in visual speech.

Second proposed approach in this research uses biometric geometrical visual features measurements from the changing shapes of the lips to make the best guess at the word being spoken. This method

uses a set of 20 landmarks around the mouth's, then calculated Mouth Aspect Ration (MAR_outer) and (MAR_inner) based on three distance, the proposed study how the points on the upper and lower lips move, as well as figuring out what is being said (the outer and the inner).

The third proposed approach using for Arabic Visual Speech Recognition (AVSR), use pre-trained deep learning algorithms Visual Geometry Group (VGG16_net), so, in the first stage: using transfer learning in lower level layers for extracted visual feature, in the second stage: which is classification stage, using new built classification layers where the model is modified by eliminating the completely connected layers that existed during model construction and replacing them with other suitable fully connected layers, with fine_tuning concept from block (5) and classification layers, which help to increase the accuracy of classification, this proposed represent M1VGG16_vsr. the M2VGG16_vsr is the same with using Data Augmentation Techniques for training phase to increase feature extraction from training data.

The system was trained and tested on local Arabic visual database, the result showed from using pre_trained deep learning algorithms VGG16 for extracting visual feature and the performance in the process of classification, as the highest classification rate which record of accuracy 0.86% found for the best implementation in M2VGG16_vsr, precision 0.86%, and recall 0.86% are all measured for identified Arabic words.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

استخراج شكل الشفاه وحركتها لتميز الكلام العربي باستخدام تقنيات التعلم العميق

اطروحة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة دكتوراه فلسفة في
علوم الحاسوب

من قبل

نغم سالم محمد الليلة

بإشراف
أ.د. خليل ابراهيم السيف

١٤٤٤ هـ ٢٠٢٣ م

الخلاصة

هدفنا في هذه البحث هو استخراج شكل الشفاه وتتبع حركتها لتخمين الكلام المرئي على مستوى كلمة في اللغة العربية, حيث ان مفهوم قراءة الشفاه يستخدم لفهم وتفسير الكلام في حالة عدم وجود حاسة السمع, وهي تقنية يعتمدها بشكل خاص الاشخاص اللذين يعانون من ضعف في مشكلة السمع لغرض تمكينهم من التواصل مع العالم الخارجي والتي بدونها لا يمكن تحقيق التواصل.

في هذا البحث اعتمدت تقنيات التعليم العميق والرؤية الحاسوبية في تفسير حركة الشفاه للكلام العربي المرئي, اما قاعدة البيانات المعتمدة تم تجميعها من خلال الفيديوهات الصامتة المسجلة للأشخاص المشاركين والبالغ عددهم عشرة اشخاص وكل مشارك ينطق تسع كلمات عربية والتي تم اعتمادها في هذا البحث, اما التسجيل كان باستخدام كاميرة الحاسوب الشخصي .

بناء وتصميم نظام تمييز الكلام العربي المرئي والذي يتكون من عدة خطوات اساسية في معالجة البيانات لغرض الحصول على دقة عالية في التمييز, فبعد عملية تجميع البيانات يتم تطبيق طريقة التدرج الموجهة التكراري Histogram Oriented Gradients(HOG) لغرض توطين منطقة الوجه واعتماد خوارزمية facial landmarks في استخراج منطقة الفم.

النظام الاول المقترح في هذا البحث لغرض تتبع حركة الشفاه بالاعتماد على شكل الشفاه ولتحقيق هذا الغرض تم استخدام Gaussian filter مرشح كاوسن ,(HSV) Hue Saturation Value , Morphological Operation and Canny filter العمليات المورفولوجية, لعزل واستخراج حافة الشفاه والتي تمتلك معلومات الكلام المرئي, ولغرض تتبع حركة الشفاه المتحركة يتم ادخال حافة الشفاه الى دالة متعددة الحدود Polynomial Function من الدرجة الثامنة ، حيث تم تكوين منحني حركة الشفاه بالاعتماد على معلمات الدالة لاشتقاق حركة الشفه اثناء الكلام المرئي.

النهج الثاني المقترح في هذه الدراسة استخدام قياسات الميزات المرئية الهندسية المستخرجة من تغيير شكل الشفاه المتحركة للحصول على افضل تخمين للكلام, ويعتمد هذا النهج على عشرون نقطة حول منطقة الفم ومن ثم حساب نسبة ابعاد الفم Mouth Aspect Ratio (MAR_out) and MAR_inner بالاعتماد على ثلاث من المسافات, بعد ذلك تم تكوين متجهه الخصائص الذي يحدد الشكل البايومتري الهندسي للشفاه المتحركة لتخمين الكلام المرئي بالاعتماد على الميزات الهندسية المستخرجة.

اما النهج الثالث في تمييز الكلام العربي المرئي, تم استخدام خوارزمية التعلم العميق Visual Geometry Group (VGG16_net) المعاد تدريبها, في المرحلة الاولى: يتم استخدام مفهوم نقل التعلم في الطبقات ذات المستويات الواطئة لغرض استخراج الميزات المرئية, اما المرحلة الثانية: والمتمثلة بمرحلة التصنيف, يتم استخدام طبقات التصنيف المعاد بناءها حيث يتم تحديث الموديل المقترح من خلال ازالة وبالكامل طبقات الاتصال الكامل واستبدالها بطبقات الاتصال الكامل المقترحة والمناسبة لتحقيق هدف البحث, اضافة الى اعتماد مفهوم fine_tuning لبلوك ذو التسلسل الخامس مع طبقات التصنيف, والذي يزيد من دقة الخوارزمية في تمييز الكلام المرئي هذا النظام المصمم في الموديل M1VGG16_vsr اما M2VGG16_vsr نفس تصميم الموديل الاول اضافة الى استخدام تقنية تضخيم البيانات Data Augumentation في مرحلة التدريب لغرض زيادة الميزات المستخرجة من بيانات التدريب.

تم تدريب النظام المقترح واختباره على قاعدة بيانات محلية وتم عرض نتائج الخوارزمية التعلم العميق VGG16 في كلا الميزات المرئية المستخرجة واداء الخوارزمية في مرحلة التمييز حيث ان اعلى نسبة تمييز تم الحصول عليها من تطبيق M2VGG16_vsr وكانت نسبة الدقة ٨٦%, precision٨٦% , recall٦٨% في تمييز الكلام العربي المرئي.