جامعة عين شمس
كلية الهندسة
قسم الحاسبات و النظم

مناهج التعلم العميق فى التعرف على الكتابة العربية

مقدم الرسالة : محمد عطية محمد رضوان

الدرجة : ماجستير العلوم الهندسية

ملخص الرسالة

التعرف على النص العربي في الوثائق المطبوعة أو المشاهد الطبيعية يعتبر من المشاكل الصعبة مقارنة مع نفس التطبيق في اللغات اللاتينية. وذلك لأن الكتابة العربية المطبوعة أكثر تعقيدا والفرق بين الأحرف يمكن أن يكون دقيق مما يزيد الخطوات اللازمة في مثل هذه النظم التى نريد بنائها. في هذا البحث تم تطوير أنظمة للتعرف على الحروف العربية لمختلف المهام. أول نظام تم بنائه يتم استخدامه للتعرف على وتدوين الحروف العربية على لوحات السيارات المسجلة فى ملفات الفيديو. يتكون النظام من خوارزمية لاستخراج اللوحة، والشبكة العصبية العميقة للتعرف على الأحرف على لوحة. تم تدريب الشبكة العصبية الاصطناعية على بيانات مصنعة ألياً واختبارها على أمثلة من العالم الحقيقي تم أستخراجها و تجهيزها يدويا. حقق نموذج التعرف على الحروف العربية ٩٠٪ من الدقة، في حين أن نموذج للأرقام العربية حقق ٩٤٪ من الدقة.

ثم نقدم نظام التعرف على الحروف العربية من وثائق النص العربي الممسوحة ضوئيا. يبدأ خط الانابيب مع وثيقة تتضمن أسطر من النص، و من ثم تقسم الى كل سطر على حدي ثم إلى كلمات باستخدام Histogram Projection Thresholding. بما أن باقى الانظمة تم تدريبها للتعرف على النص بحجم 18pt، قمنا ببناء نموذج شبكة عصبية لتوقع حجم مدخلا عبارة عن صورة كلمة، وبالتالي يمكننا أن نغير حجم الكلمة الى حجم 18pt.

نموذج شبكة عصبية عميقة متعددة المدخلات الفرعية تم استخدامه لتقسيم الكلمة إلى أحرف. ثم نموذجا اخر للتعرف على الحروف يستخدم في النهاية ليكون الناتج النهائي.

أجريت تجارب للشبكة العصبية لتقسيم الكلمة الى حروف لنحقق دقة ٩٨.٩٪. تم اختبار المنظومة الكاملة للتعرف على الكتابة العربية على جزء من مجموعة البيانات APTI لتحقق دقة ٩٤.٣٨٪.

وينقسم البحث إلى ستة فصول بما في ذلك قوائم المحتويات وقائمة المراجع.

الفصل الاول

مقدمة في مكونات وأنظمة التعرف على الحروف العربية مثل: تحليل المخططات، تجهيزها، والتعرف عليها، و مرحلة ما بعد التعرف.

الفصل الثاني

دراسة لمفاهيم مختلفة فى الشبكات العصبية والتعلم العميق التي استخدمناها لبناء نماذج لدينا.

الفصل الثالث

عرض بحثنا فى النظام التلقائي للتعرف على لوحات السيارات العربية.

الفصل الرابع

يقدم هذا الفصل الجزء الأول من نظامنا للتعرف على الوثائق العربية. نصف خوارزميات تجزئة الوثيقة الى خطوط و كلمات، ومن ثم نقدم نموذج التنبؤ بحجم الخط.

الفصل الخامس

يشرح شبكة عصبية عميقة متعددة المدخلات الفرعية لتجزئة الكلمات، ومن ثم نموذج للتعرف على الحروف العربية.

الفصل السادس

لخاتمة والعمل المستقبلي المخطط له.

المشرفون:

ا. د./ حازم محمود عباس

د./ محمود ابراهيم خليل

Ain Shams University
Faculty of Engineering
Computer and Systems
Department

Deep Learning Approaches in Arabic OCR

Submitted by: Mohamed Atia Mohamed Radwan

Degree: Master of Science

Thesis Summary

Recognition of Arabic text in printed documents or natural scenes is a hard problem compared to the same application on Latin languages. Since Arabic typography is more complicated and the difference between characters can be very subtle there becomes more steps needed in such systems we want to build. In this thesis we develop Arabic characters' recognition systems for different tasks. The first recognition pipeline is used to recognize and transcribe Arabic licenses car plates from a video stream. The pipeline consists of a localization algorithm for extracting the license plate, and a deep neural network for recognizing characters on the plate. The neural network was trained on synthetic data and tested on a real world example manually annotated. The model for Arabic character recognition achieved 90% accuracy, while the model for Arabic numbers recognition achieved 94% accuracy. We then introduce an Arabic OCR system for recognizing Arabic text in scanned documents. The pipeline starts with a document containing lines of text, and segments them into each line alone then into sub-words using histogram projection thresholding. Since the remaining sub-systems of the pipeline are trained to recognize text of default size 18pt, we built a neural network model for predicting the size of an input sub-word, hence we can afterwards normalize this sub-word into the 18p default size. A multichannel deep neural network model is built to segment input sub-word into characters. Then a model for recognizing characters is finally used to have the final output.

Experiments for the segmentation neural network were carried on a computer generated test data. The model for segmenting one font had a 98.9% accuracy. The whole Arabic OCR pipeline was tested on a subset of the APTI dataset and scored a 94.38% test accuracy.

The thesis is divided into six chapters including lists of contents and a list of references.

Chapter 1

Introduction to the components of an Arabic character's recognition systems like: Layout analysis, Preprocessing, Recognition, and Post-processing are defined and the different steps in each component are discussed.

Chapter 2

Survey for the different concepts of Neural Networks and deep learning that we used for building our models.

Chapter 3

Presentation of the research for automatic Arabic car license plate recognition system.

Chapter 4

This chapter introduces the first part of our Arabic OCR which is the document and lines segmentation algorithms, and then we present the font size prediction model.

Chapter 5

Covers our novel multi-channel neural network for sub-words segmentation, and then the conv-net for Arabic characters' recognition.

Chapter 6

It covers the Conclusion and our planned future work.

Supervisors:

Prof. Dr. Hazem Mahmoud Abbas

Dr. Mahmoud Ibrahim Khalil

AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer Engineering and Systems

Deep Learning Approaches in Arabic OCR

A Thesis submitted in partial fulfillment of the requirements of
Master of Science
Computer Engineering and Systems

by
Mohamed Atia Mohamed Radwan
Bachelor of Science
Computer Engineering and Systems
Faculty of Engineering, Mansoura University, 2010

Supervised By
Prof. Hazem M. Abbas
Dr. Mahmoud I. Khalil

Cairo, 2017

AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer Engineering and Systems

Deep Learning Approaches in Arabic OCR

by

Mohamed Atia Mohamed Radwan

Bachelor of Science in Computer Engineering
Computer Engineering and Systems
Faculty of Engineering, Mansoura University, 2010

Examiners' Committee

Name and affiliation	**Signature**
Prof. Mohsen A. Rashwan Faculty of Engineering, Cairo University.	.....
Prof. Hussain I. Shaheen Faculty of Engineering, Ain Shams University.	.....
Prof. Hazem M. Abbas Faculty of Engineering, Ain Shams University.	.....
Dr. Mahmoud I. Khalil Faculty of Engineering, Ain Shams University.	.....

Date:08 April 2017

Statement

This thesis is submitted as a partial fulfillment of Master of Science in Computer Engineering and Systems, Faculty of Engineering, Ain shams University. The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Mohamed Atia Mohamed Radwan

Signature

.....

Date: 08 April 2017

Researcher Data

Name: Mohamed Atia Mohamed Radwan
Date of Birth: 05/04/1988
Place of Birth: Mansoura, Egypt
Last academic degree: Bachelor of Science
Field of specialization: Computer Engineering and Systems
University issued the degree : Mansoura University
Date of issued degree : 2010
Current job : Data Scientist

Thesis Summary

Recognition of Arabic text in printed documents or natural scenes is a hard problem compared to the same application on Latin languages. Since Arabic typography is more complicated and the difference between characters can be very subtle there becomes more steps needed in such systems we want to build. In this thesis we develop Arabic characters recognition systems for different tasks. The first recognition pipeline is used to recognize and transcribe Arabic licenses car plates from a video stream. The pipeline consists of a localization algorithm for extracting the license plate, and a deep neural network for recognizing characters on the plate. The neural network was trained on synthetic data and tested on a real world example manually annotated. The model for Arabic character recognition achieved 90% accuracy, while the model for Arabic numbers recognition achieved 94% accuracy. We then introduce an Arabic OCR system for recognizing Arabic text in scanned documents. The pipeline starts with a document containing lines of text, and segments them into each line alone then into sub-words using histogram projection thresholding. Since the remaining sub-systems of the pipeline are trained to recognize text of default size 18pt, we built a neural network model for predicting the size of an input sub-word, hence we can afterwards normalize this sub-word into the 18pt default size. A multichannel deep neural network model is built to segment input sub-word into characters. Then a model for recognizing characters is finally used to have the final output.

Experiments for the segmentation neural network were carried on a computer generated test data. The model for segmenting one font had a 98.9% accuracy. The whole Arabic OCR pipeline was tested on a subset of the APTI dataset and scored a 94.38% test accuracy. The thesis is divided into six chapters including lists of contents and a list of references.

Chapter 1 Introduction to the components of an Arabic characters recognition systems like: Layout analysis, Preprocessing, Recognition, and Post-processing are defined and the different steps in each component are discussed.

Chapter 2 Survey for the different concepts of Neural Networks and deep learning that we used for building our models. Chapter 3 Presentation of the research for automatic Arabic car license plate recognition system.

Chapter 3 This chapter introduces the first part of our Arabic OCR which is the document and lines segmentation algorithms, and then we present the font size prediction model.

Chapter 5 Covers our novel multi-channel neural network for sub-words segmentation, and then the conv-net for Arabic characters recognition.

Chapter 6 It covers the Conclusion and our planned future work.