IMAGE AND VIDEO BASED EMOTION RECOGNITION USING DEEP LEARNING

BY

ARSELAN ASHRAF

A dissertation submitted in fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering)

Kulliyyah of Engineering
International Islamic University Malaysia

MARCH 2021

ABSTRACT

Emotion recognition utilizing pictures, videos, or speech as input is considered an intriguing issue in the research field over certain years. The introduction of deep learning procedures like the Convolutional Neural Networks (CNN) has made emotion recognition achieve promising outcomes. Since human facial appearances are considered vital in understanding one's feelings, many research studies have been carried out in this field. However, it still lacks in developing a visual-based emotion recognition model with good accuracy and uncertainty in determining influencing features, type, the number of emotions under consideration, and algorithms. This research is carried out to develop an image and video-based emotion recognition model using CNN for automatic feature extraction and classification. The optimum CNN configuration was found to be having three convolutional layers with max-pooling attached to each layer. The third convolutional layer was followed by a batch normalization layer connected with two fully connected layers. This CNN configuration was selected because it minimized the risk of overfitting along with produced a normalized output. Five emotions are considered for recognition: angry, happy, neutral, sad, and surprised, to compare with previous algorithms. The construction of the emotion recognition model is carried out on two datasets: an image dataset, namely "Warsaw Set of Emotional Facial Expression Pictures (WSEFEP)" and a video dataset, namely "Amsterdam Dynamic Facial Expression Set – Bath Intensity Variations (ADFES-BIV)." Different pre-processing steps have been carried over data samples, followed by the popular and efficient Viola-Jones algorithm for face detection. CNN has been used for feature extraction and classification. Evaluating results using confusion matrix, accuracy, F1-score, precision, and recall shows that video-based datasets obtained more promising results than image-based datasets. The recognition accuracy, F1 score, precision, and recall for the video dataset came out to be 99.38%, 99.22%, 99.4%, 99.38, and that of the image dataset came out to be 83.33%, 79.1%, 84.46%, 80%, respectively. The proposed algorithm has been benchmarked with two other CNN-based algorithms, and the accuracy performs better around 5.33% and 3.33%, respectively, for the image dataset, while 4.38% for the video dataset. The outcome of this research provides the productivity and usability of the proposed system in visual-based emotion recognition.

خلاصة البحث

يعتبر التعرف على المشاعر باستخدام الصور أو مقاطع الفيديو أو الكلام كمدخلات ، قضية مثيرة للاهتمام في مجال البحث على مر السنين. لقد حقق إدخال إجراءات التعلم العميق مثل الشبكات العصبية التلافيفية والتعرف على المشاعر نتائج واعدة في هذا المجال. نظرًا لأن ملامح الوجه البشرية تمثل سمات مهمة في فهم مشاعر المرء. تم إجراء العديد من الأبحاث في هذا المجال ، لكن هذه الأبحاث لا تزال تفتقر إلى تطوير نموذج مرئي للتعرف على المشاعر بدقة واعدة، بالإضافة إلى عدم اليقين في تحديد السمات المؤثرة ، ونوع وعدد المشاعر قيد الدراسة ، والخوارزميات. تم إجراء هذا البحث لتطوير نموذج يعتمد على الصورة والفيديو للتعرف على المشاعر ، وذلك باستخدام CNN لاستخراج الميزات وتصنيفها تلقائيًا. تم استنتاج أن التكوين الأمثل لـ CNN له ثلاث طبقات تلافيفية ، مع أقصى تجمع مرتبط بكل طبقة. عقبت الطبقة التلافيفية الثالثة طبقة تسوية حزمة متصلة بطبقتين متصلتين بالكامل.تم اختيار تكوين CNN هذا لأنه يقلل من مخاطر التركيب الزائد مع الناتج الطبيعي. تم أخذ خمسة مشاعر في الاعتبار للتعرف عليها: غاضب ، سعيد ، محايد ، حزين ، ومندهش، وذلك للمقارنة مع الخوارزميات السابقة. تم تنفيذ بناء نموذج التعرف على المشاعر على مجموعتي بيانات: مجموعة بيانات للصور وهي "مجموعة وارسو من صور تعبيرات الوجه العاطفية (WSEFEP)" ومجموعة بيانات فيديو تسمى "مجموعة تعبير الوجه الديناميكي بأمستردام - تنويعات كثافة الاستحمام (ADFES-BIV) ". تم تنفيذ خطوات معالجة مسبقة مختلفة على عينات البيانات متبوعة بخوارزمية فيولا جونز الشائعة والفعالة لاكتشاف الوجه. تم استخدام CNN لاستخراج الميزات والتصنيف. تظهر نتائج التقييم عند استخدام مصفوفة الارتباك ودقة التعرف ودرجة F1 والدقة والاستدعاء أن مجموعة البيانات المستندة إلى الفيديو حصلت على نتائج واعدة أكثر، مقارنة بمجموعة البيانات القائمة على الصور. بلغت دقة التعرف ودرجة F1 والدقة والاستدعاء لمجموعة بيانات الفيديو 99.38٪، 99.22٪، 99.4٪، و99،38٪، على التوالي، بينما كانت تلك الخاصة بمجموعة بيانات الصور 83.33٪، 79.1٪، 84.46٪ و 80٪، على التوالي. تم اختبار الخوارزمية المقترحة مع خوارزميتين أخريين تعتمدان على CNN ، حيث أظهرت أداءً أفضل من حيث الدقة حوالي 5.33٪ و 3.33٪ على التوالي لمجموعة بيانات الصورة ، بينما أظهرت تحسنًا بنسبة 4.38٪. لمجموعة بيانات الفيديو. توفر نتائج هذا البحث إنتاجية وإمكانية استخدام النظام المقترح في التعرف على المشاعر المرئية.

APPROVAL PAGE

I certify that I have supervised and read this study and that in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Master of Science (Computer and Information Engineering)

…………………………………..
Teddy Surya Gunawan
Supervisor

…………………………………..
Farah Diyana Abdul Rahman
Co-Supervisor

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Master of Science (Computer and Information Engineering)

…………………………………..
Khairul Azami Sidek
Internal Examiner

…………………………………..
Hasmah Mansor
Internal Examiner

This dissertation was submitted to the Department of Electrical and Computer Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering)

…………………………………..
Mohamed Hadi Habaebi
Head, Department of Electrical
and Computer Engineering

This dissertation was submitted to the Kulliyyah of Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering)

…………………………………..
Sany Izan Ihsan
Dean, Kulliyyah of Engineering

DECLARATION

I hereby declare that this dissertation is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Arselan Ashraf

Signature: Date: 15/03/2021

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

IMAGE AND VIDEO BASED EMOTION RECOGNITION USING DEEP LEARNING

I declare that the copyright holders of this dissertation are jointly owned by the student and IIUM.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.

3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Arselan Ashraf

15/03/2021

....

Signature Date

ACKNOWLEDGEMENTS

Firstly, I would like to thank Almighty Allah for blessing me with good health and composure for this research. It is my utmost pleasure to dedicate this work to my dear parents and my family, who granted me the gift of their unwavering belief in my ability to accomplish this goal: thank you for your support and patience.

I wish to express my appreciation and thanks to those who provided their time, effort, and support for this project.

Finally, a special thanks to my supervisor Prof. Dr. Teddy Surya Gunawan and co-supervisor Dr. Farah Diyana for their continuous support, encouragement, and leadership, and for that, I will be forever grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$Cdf(min)$	Base zero estimations of total conveyance work
$f_1 f_2 f_3$	Features
$f(x)$	Generic Classifier
L	Gray levels
$H(v)$	Histogram balance
$\alpha_1 \alpha_2 \alpha_3$	Individual Loads of Features
$M \times N$	Number of pixels

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Networks
CNN	Convolutional Neural Networks
DL	Deep Learning
DNN	Deep Neural Network
KNN	K-Nearest Neighbors
RNN	Recurrent Neural Network
SVM	Support Vector Machine

CHAPTER ONE
INTRODUCTION

1.1 BACKGROUND OF THE STUDY

In recent times, emotion recognition has evolved as one of the main highlights in the domain of artificial intelligence. The gigantic expansion in the improvement of modern human-computer collaboration advancements has additionally helped the movement of advancements pertaining to this sphere. Facial activities pass on the feelings which thusly pass on an individual's character, state of mind, and expectations. Feelings generally rely on the facial highlights of a person alongside the voice. Be that as it may, there are some different highlights too, specifically physiological highlights, social highlights, actual highlights of the body, and some more. Several works have been done to recognize emotions with more exactness and accuracy. The objective of feeling acknowledgment can be accomplished by utilizing visual-based methods or sound-based procedures. AI has changed the field of computer-human collaboration and gives many Machine Learning methods to arrive at our point. Many machine learning techniques are present to perceive the feeling, however, this research will focus on image and video based feeling acknowledgment utilizing DL. Image and Video-based feeling acknowledgment is multidisciplinary and incorporates fields like brain science, emotional figuring, and human-PC connection. Facial expressions consist of 55% of the emotion of an individual (C.-H. Wu, J.-C. Lin and W.-L. Wei, 2014).

To create a well-fitted model for image and video based feeling acknowledgment, an appropriate feature casings of the facial appearance must be available. Rather than utilizing ordinary methods, deep learning gives an assortment regarding precision, learning rate, forecast, and so on. CNN is among one of the deep learning strategies which have offered help and stage for examining visual symbolism. Convolution is the basic utilization of a channel to information that result in an activation. Repeated utilization of a comparative channel to an info achieves a guide of establishments called an element map, indicating the regions and nature of a perceived component in contribution, for instance, an image. The improvement of convolution neural frameworks is the ability to subsequently pick up capability with a huge number of channels in equivalent unequivocal to a preparation dataset under the necessities of a specific insightful showing issue, for instance, picture portrayal. The result is significantly clear features that can be recognized anyplace on input pictures. Deep learning has made incredible progress in perceiving the feelings, and CNN is the notable deep learning strategy that has accomplished a wonderful exhibition in picture preparation. There has been a lot of work in visual pattern acknowledgment for facial emotion recognition, similarly as in signal preparing for sound-based acknowledgment of sentiments. Moreover, there are a number of multimodal approaches joining these prompts (Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang, 2009). From past decades, there has been a rapid rise in research in computer vision on facial expression analysis (V. P.c. and N. K.r., 2015). Inspired by deep learning, this research aims to formalize an image and video based emotion recognition model.

1.2 PROBLEM STATEMENT

Facial expressions are the main features of the emotions of an individual. Human facial emotions are the fundamental ways for conveying information among people. Exchange of emotions can happen during conversation, resulting in change in the facial expressions. Although much research has been conducted in this sphere, however the methods that are present are lacking performance in terms accuracy. The methods with better accuracy (in 80 %) are facing low performance in terms of precision, recall and F1 score. Majority of the emotion recognition models are evaluated using passive audio or image-based datasets. With the inclusion of more emotions the performance parameters of the model tend to decrease. These problems provided an encouragement to conduct this research.

1.3 RESEARCH OBJECTIVES

The prime objective of this research is to extract and analyze visual features from the image and video files using MATLAB, then classifying those features using CNN. The objectives are listed as under:

1- To investigate and analyse various image and video databases and select two standard datasets; image based and video based.

2- To design an integrated image and video based facial emotion recognition model using convolutional neural networks.

3- To evaluate the performance parameters of the proposed recognition model in terms of accuracy, precision, recall, F1-score and confusion matrix.

1.4 RESEARCH METHODOLOGY

The basic architecture for developing an image and video based emotion recognition model using DL is shown in Figure 1.1.

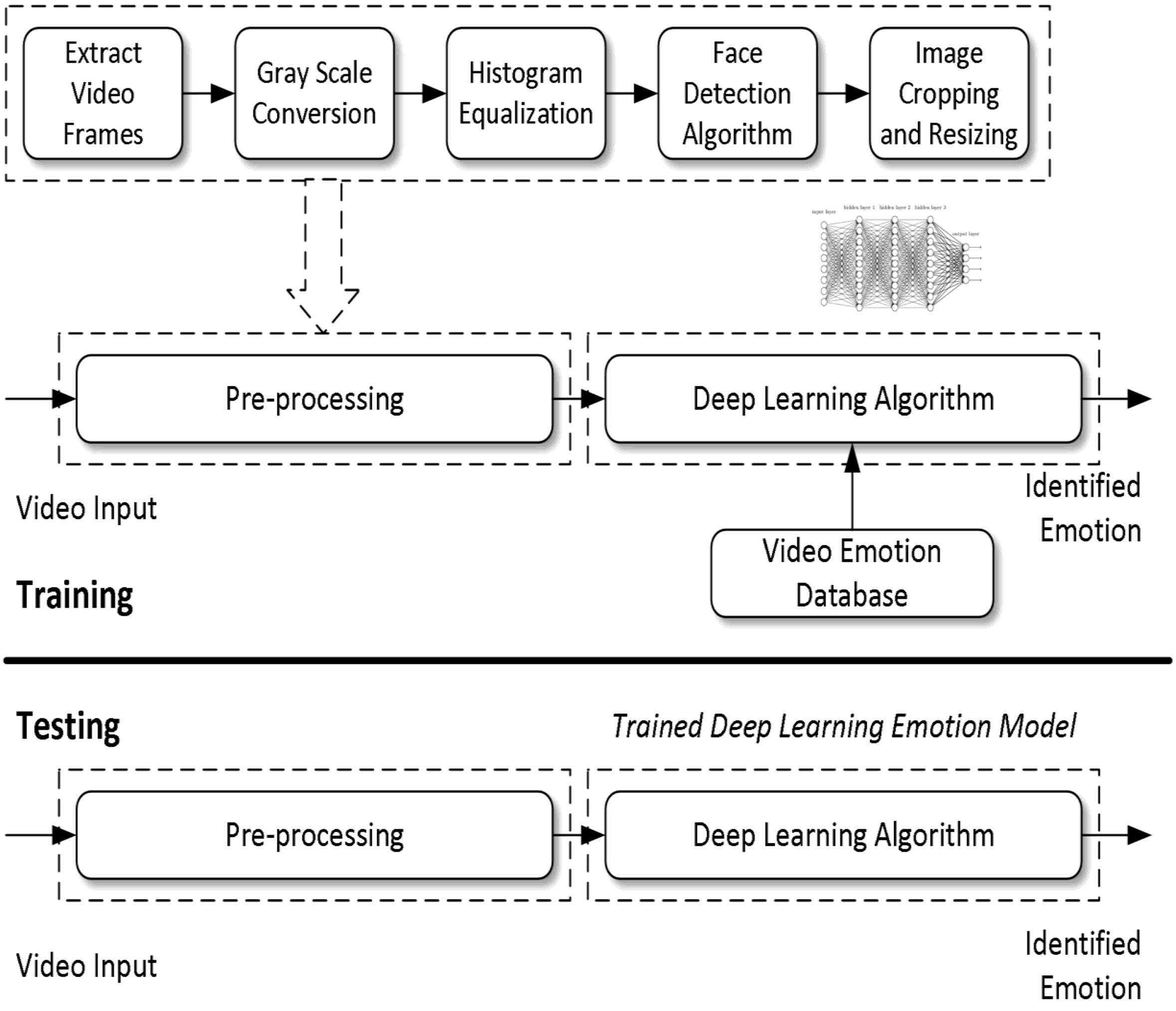

Figure 1.1 Architectural Diagram

As in image/video-based emotion recognition, the input visual samples are processed, which includes several preprocessing steps, also features are extracted from the face. Since facial features are important for emotion recognition using images and videos, these features are then subjected to the training algorithm for the development of a well fitted model.

1.5 RESEARCH SCOPE

This research aims to create an image and video based emotion recognition model using convolutional neural networks. Two databases are used in this project one image and another video. The technique of Convolutional Neural Networks is considered for model training and testing. This work is focused upon using images or video as input. Apart from that, no other source will be considered.

1.6 THESIS ORGANIZATION

The flow of this dissertation is categorized as follows. Chapter 2 includes a literature review and discusses research conducted relating to image/video-based emotion recognition and DL. Chapter 3 includes the methodology and implementation of the research. The results and discussion are elaborated in Chapter 4. Finally, Chapter 5 presents the conclusion, benchmarking, and future recommendations.

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

Emotion recognition is one of the trending hot topics in the sphere of research. Facial expressions are the significant implications of one's emotions. Therefore, to determine the mood of an individual, facial expressions are to be recognized accurately. With the inclusion of Artificial Intelligence techniques in the sphere of emotion recognition, there has been a promising rise in better results and more accurate performance parameters. According to Lens Organization (http://lens.org) the rise in the interests of various researchers in this field has tremendously grown over the time. This growth can be clearly analyzed from the Figure 2.1.

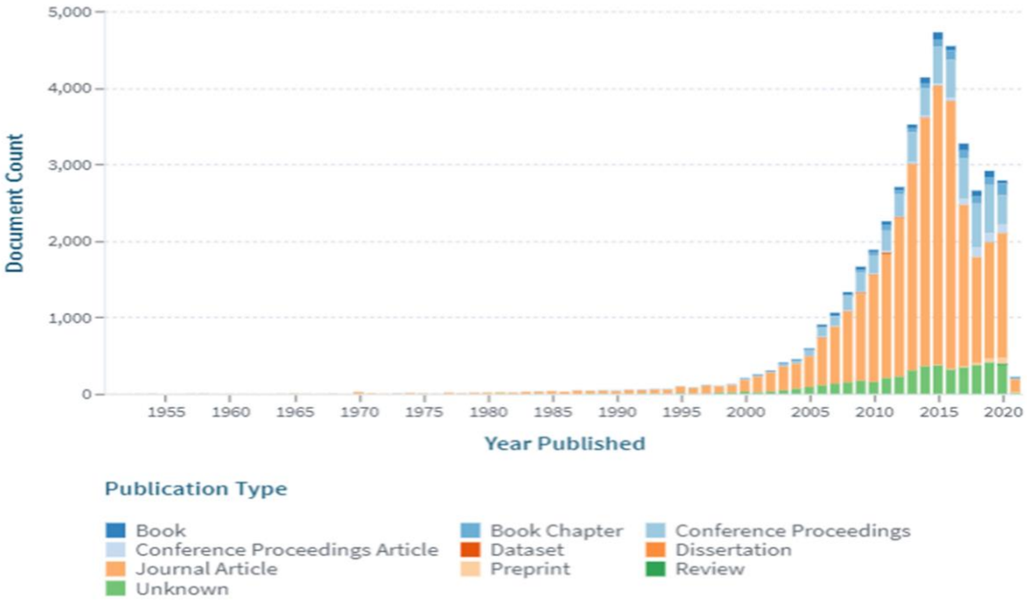

Figure 2.1 Scholarly works in the sphere of emotion recognition

According to (Y.Cai, W.Zheng, T. Zhang, Q. Li, Z. Cui, and J. Ye, 2016), they developed a Video ER model using CNN-RNN and C3D (type of CNN containing 8 layers of convolution, 5 layers of max-pooling, 2 fully connected layers , subsequently a softmax layer) Hybrid Networks by extracting and aligning all facial frames present in the video and then transforming them with respect to the facial vital points. In case of falsely detected faces, CNN based face filtering was performed. In case of RNN training, sixteen facial features were arbitrarily selected. For each video clip sixteen facial frames were given as input to the C3D network, which proved 59.02% accurate for the testing set. According to (Jirayucharoensak, S., Pan-Ngum, S., &Israsena, P., 2014), EEG based emotion recognition system is implemented with a stack of three auto encoders with two softmax layers. Their system performed emotion recognition by estimation valence and arousal states separately. The technique used in this model was DLN utilizing unsupervised pertaining technique with greedy layer wise training.

According to (T. S. Wingenbach, C. Ashwin, and M. Brosnan, 2016), they made and endorsed a bunch of video recordings portraying three levels of facial emotion intensities, from low to high power. The samples were adjusted from the Amsterdam Dynamic Facial Expression Set Bath Intensity Variations dataset, completing a facial inclination acknowledgment task, which recollected six basic emotions in extension to pride, disgrace and contempt, which were imparted at three unique forces of appearance and neutrality. Precision rates over the opportunity level of reacting were found for all feeling classifications, delivering general crude hit pace of 69% for ADFES-BIV. In, (Sonmez, 2018) tested the grouping explore run on the ADFES-BIV dataset. The proposed programmed framework utilizes the scanty portrayal-based classifier and arrives at the top execution of 80% by considering the worldly data characteristically present in the videos. According to (Fan, Y., Lam, J. C.

K., & Li, V. O. K., 2018), in video based emotion recognition using deeply supervised CNN the objective is to enhance the component guide of each layer, by joining the associations over the side-yield layers. To this end, they embrace de-convolution methods in the up sampling activity, which can take the contribution of a discretionary size and produce size yield correspondingly.

One of the significant drivers of research right now been the emotion recognition in the wild challenges, which presented and built up an out of research facility dataset namely acted facial expressions in the wild, gathered from recordings that copy reality. The EmotiW Challenge, which began in 2013, intends to beat the difficulties of information assortment, comment, and estimation for multimodal feeling acknowledgement in nature. The test utilizes the AFEW corpus, which mostly comprises of motion picture extracts with uncontrolled conditions (Abhinav Dhall, Roland Goecke, Jyoti Joshi, MichaelWagner, Tom Gedeon, 2013). (Reeshad Khan & Omar Sharif, 2017) in their literature review on emotion recognition using various methods, proposed utilizing EEG and various media signal yields the ideal outcomes. They accepted Long Short-Term Memory Network Recurrent Neural Network (LSTM-RNN) is the ideal approach to deal with multimodalities. So, their proposition was centered on emotion recognition by EEG and broad media signal utilizing LSTM-RNN. This kind of research has been done previously. But their test was to improve the model where it will be prepared by EEG and varying media information simultaneously and will make a connection between this information wherein, on the off chance that one sort of information isn't accessible in a circumstance, the model could, in any case, produce the outcome, finding the connection inside the information. Some more scholarly works are present in Table 2.1.