Scientific Computing Department
Faculty of Computer and Information Sciences
Ain Shams University

Action Recognition in videos using Deep Learning

Thesis submitted as a partial fulfillment of the requirements for the degree of
Master of Science in Computer and Information Sciences

By
Bassel Safwat Chawky Hakim

Teaching Assistant at Scientific Computing Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Supervised by:

Prof. Dr. Howida Abdel Fattah Shedeed
Professor in Scientific Computing Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Mohammed Abd El-Rahman Marey
Assistant Professor in Scientific Computing Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Ahmed Samir
Ph. D. in Scientific Computing,
Faculty of Computer and Information Sciences,
Ain Shams University

June – 2019
Cairo

Acknowledgment

First of all, I would to like to thank GOD for his endless blessings, for giving me the power and strength to complete this work and for surrounding me supportive people.

I would like also to thank my Professor Dr. Howida. Her office door was always open when I ran into trouble or had a question about my research or writing. She consistently allowed our researches to be my own work but steered me in the right direction whenever she thought I needed it. Also, I would like to express my deep appreciations for the scientific vision of Dr. Mohammed Marey and his support through all the meetings that lasted till late evenings besides his continuous pushes to reach my extremes and to achieve several successes. Next, I would like to thank Dr. Ahmed Samir Elons not only for his scientific vision and guidance but also for his patience during my early years in my masters.

Special thanks for my family who always believed in me and supported me through my toughest times. I would like to explicitly thank my mother for her permanent trust and acceptance for the way I am, pushing me to always do my best and to seek quality in my work. Thank you again for all the efforts you put to allow me achieving my goals and targets and your endless support.

My friends who always look at me as a man with potential, thank you for your trust and for existing in my life. I appreciate those time when you kept asking me to pursuit my dreams and to finish my masters successfully.

Special thanks for the theatre team for their efforts to record the dataset and support for the research. Last but not least, I would like to thank my professors, colleagues and students who kept encouraging me. Thank You!

Bassel Safwat

Abstract

Human action recognition for a given video is a difficult problem containing many challenges ranging from partial occlusion to variations in the action speeds and viewpoints. However, this problem is at the very core of various systems like abnormal behavior detection, action localization and/or online video analysis, and that is why it is a very important problem which got the attention from many researchers in the past decades and till now.

Despite the number of studies in the literature, the action recognition problem remains a difficult problem. Traditional approaches require a lot of efforts to find the best combination of features not only to represent the action in a compact form but also to handling the so many existing challenges. Recent methodologies relied on deep learning models to learn and extract good representation from the datasets. Although the deep learning-based models requires a large training datasets and costly training time, this type of models illustrated advances on several action recognition dataset. Moreover, several techniques like transfer learning allowed faster convergence of the accuracy by pretraining the model.

At first, a novel ranking and listing for 14 action recognition datasets is set. The ranking is based on the number of challenges each dataset covers. Therefore, the higher the dataset's rank, the more realistic the dataset is, indicating its ability to provide a realistic measurement for the models. Based on these datasets, a comparison between the advances in traditional approaches and deep-learning based models is illustrated. Based on this deep survey, deep learning baselines models, namely two stream convolutional neural network and 3D convolutional neural networks, are described. These

baseline models are almost used by all the other studies including state of the art models.

Secondly, a novel action recognition benchmark has been set and used to study several action recognition challenges like the change in viewpoints and shaky videos effects using the baseline models described. Moreover, it is used to study the overfitting problem of the deep learning models with potential solutions.

Finally, two main techniques are suggested where both can be integrated by the several existing action recognition models in order to improve the accuracy. These techniques leverage the video temporal dimension to learn several features representing varying temporal lengths. The first technique is called **S**ingle **T**emporal **R**esolution **S**ingle **M**odel (STR-SM) which suggests training the desired model on one specific temporal resolution of the video. The temporal resolution is defined as the number of frames for a specific amount of time. Therefore, a low temporal resolution means that a small number of frames is used to represent the action while for a high temporal resolution, a large number of frames is used. Therefore, a good model that uses the STR-SM technique uses a temporal resolution that is low enough to represent long temporal duration but also, high enough to capture the motion details. Such technique is faster when compared to traditional approaches as it tackles long temporal range at once with better accuracy as it covers more information. On The second technique is called **M**ulti **T**emporal **R**esolution **M**ulti **M**odel (MTR-MM) which tackles the problem of varying action speeds in a novel way. Applying the MTR-MM technique on the desired model requires building several STR model versions, each trained on a specific temporal resolution with a late fusion. This leverage the different existing information in each temporal resolution leading to a better and improved accuracy. Additionally, the STR-SM and the MTR-MM techniques are applied on 3D Convolutional Neural Network model and have improvements over the traditional training approach of 3.63% and 6% video-wise accuracy respectively.

List of Publications

1. Chawky, Bassel S., A. S. Elons, A. Ali, and Howida A. Shedeed. "A Study of Action Recognition Problems: Dataset and Architectures Perspectives." In **Advances in Soft Computing and Machine Learning in Image Processing**, pp. 409-442. Springer, Cham, 2018.
2. Chawky, Bassel S., A. S. Elons, A. Ali, and Howida A. Shedeed. "OA18: Deep Learning for Office Actions Analysis. In **2019 9th IEEE International Conference on Intelligent Computing and Information Systems (ICICIS)**, (waiting acceptance)
3. Chawky, Bassel S., Mohammed Marey, and Howida A. Shedeed. "Multi-Temporal-Resolution Technique for Action Recognition using C3D: Experimental Study." In **2018 13th International Conference on Computer Engineering and Systems (ICCES)**, pp. 404-410. IEEE, 2018.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	**Stands for**
CNN	**C**onvolutional **N**eural **N**etwork
CRF	**C**onditional **R**andom **F**ield
FV	**F**isher **V**ectors
HOF	**H**istogram of **O**ptical **F**low
HOG	**H**istogram of **O**riented **G**radients
HVS	**H**uman **V**isual **S**ystem
LDA	**L**atent **D**irichlet **A**llocation
LSTM	**L**ong-**S**hort-**T**erm **M**emory
MBH	**M**otion **B**oundary **H**istogram
MKL	**M**ultiple **K**ernel **L**earning
ReLU	**Re**ctified **L**inear **U**nits
SFA	**S**low **F**eature **A**nalysis
SFV	**S**tacked **F**isher **V**ectors
SGD	**S**tochastic **G**radient **D**escent
STR-SM	**S**ingle **T**emporal **R**esolution **S**ingle **M**odel
TDD	**T**rajectory-pooled **D**eep-convolution Descriptors
TS	**T**rajectory **S**hape

Chapter 1

Introduction

Chapter 1. Introduction

1.1 Action Recognition and Deep Learning Overview

Action recognition is the task of identifying the kind of action presented in the video. The action occurring in the videos vary in complexity, where simple actions are called gestures and complex action is called activity. Therefore, it is important to provide a clear definition for the different action complexities.

Gesture: This is the basic motion performed by the human body parts that have some meaning. This includes hands or legs movements like waving hands or walking, facial expressions like closing/opening eyes and lips movements, and head shaking. This is considered as atomic action as it has the lowest complexity and the shortest amount of time when compared with other action complexities.

Action: There is no standard definition for an ‘action’, however, it is commonly understood as the movements (more than one gesture) performed by a person that involves an interaction with another person or object. In this context, shaking hands, swimming, and kicking a ball are good examples on actions. They usually require orders of few seconds and might last for minutes.

Activity: It is a set of actions that occurs either simultaneous or in sequence. Therefore, it is the most complex type of actions and is also called **event**. Examples for different activities: people protesting, a team game or a group meeting. As might be guessed, these activities usually last for long time interval.

This thesis focuses on improving generic deep learning systems for the 'gesture' and 'action'. These actions complexity covers the basic human actions in addition to interaction with another person and or objects.

Following is a discussion of few important applications for action recognition task to illustrate the different areas that are affected by any advances in action recognition.

A. Video Retrieval [1]: As the number of videos on the internet is growing, there is a tremendous need to automatically understand the content of the video instead of relying solely on the video tags (i.e. title and description), which might be misleading, to improve the search results quality for the end users.

B. Automated Surveillance [2]: The surveillance in large manufactures and governmental institutes requires security cameras that covers the whole place to record the daily activities occurring. There is usually a security person monitoring these videos to notify for any abnormal behaviors and take the required decision. Automating this task would benefit in improving the accuracy and reducing the costs for workers performing a tedious task.

C. Human-Computer Interaction [3]: Many modern games provide sensors like web-cameras, kinetic or Virtual Reality (VR) controllers to capture the human action and provide a better entertaining gaming experience. The quality of understanding the human action results in a better user experience.

D. Video Description and Summarization [4]: Similar to video retrieval, generating a better video description is another task that relies on understanding the video content. Increasing such system with actions