Human Action and Interaction Recognition in surveillance videos

Master of Science Thesis

By
DELIANG WU

Utrecht University

Department of Information and Computing Science
UNIVERSITY UTRECHT
Supervisor: dr. ir. R.W. Poppe

Wuhan University

School of Remote Sensing and Information Engineering
WUHAN UNIVERSITY
Supervisor: Prof. dr. Zhenzhong Chen

A dissertation submitted to the University of Utrecht in accordance with the requirements of the degree of MASTER OF SCIENCE in Artificial Intelligence .

JANUARY 2017

Word count: ******

Abstract

Here goes the abstract

Dedication and acknowledgements

Here goes the dedication.

TABLE OF CONTENTS

List of Tables

Table **Page**

List of Figures

Figure **Page**

CHAPTER 1

INTRODUCTION

The technique of automatic video content analysis(VCA) is one of the most important branch of Computer Vision and Artificial Intelligence. With this technique, machine can recognize objects, human activities and events in videos. Thus, it can be widely used in many domains, such as human-computer interaction, video classification, entertainment, self-driving, public safety and security, home automation, etc. Action and interaction recognition/detection is one of the most common use of VCA which focus on human activities analysis, including action recognition/detection of one person or interaction recognition/detection across two or more people.

1.1 Background

Talking about the technique of action and interaction recognition/detection in videos, we better start with that in the simpler single image case. Usually, the action and interaction recognition in a single image includes human detection, single person activity recognition and higher level group people activity recognition. Though the researchers have done a lots of excellent works on this topic, but it is still far away can be called solved. It is still a very challenging and hot researching area. The goal of action and interaction recognition in video is the same as that in the single image but the former is more challenging due to the introducing of temporal information.

The main challenges of action and interaction recognition and detections in real scenes mainly include:

1. Various camera viewpoints, the videos which will be analysed could be taken from different viewpoints which have never been seen in training data.

2. Complex background, the background of the interested action and interaction could be various and even totally different.

3. Various in lighting, the same class of action and interaction could be taken in totally different lighting conditions.

4. Partly occlusion between people.

5. Usually hundreds or more frames in a single videos, and therefore there are many irrelevant frames which would confuse the analysis.

6. It is hard to get a decent performance on a small training set, but training on a large scale video data-set is highly complex and computing resources consuming.

The researches in the area of video action and interaction recognition/detection can be mainly divided into two directions: hand-crafted feature descriptions and deep learning feature descriptions in the recent years. In the direction of hand-craft feature descriptions, many popular methods in image analysis like SIFT, HOG are extended to 3D-SIFT[5],[2] and 3D-HOG[4] for action recognition in videos. Among all works in this direction, the method improved Dense Trajectories(iDT)[8] is the state-of-art work prosed by Wang et al.

As mentioned before about the big challenges of action and interaction recognition/detection in video, such as various in camera viewpoints, background, lighting, camera motion, etc. which are hard for hand-crafted feature descriptor. So, more and more researchers in this area began to use deep learning as the feature descriptor. There are many works in the earlier stage which just simply considered videos as set of single frames and averaged the results over all frames as the overall results[3]. The performance of such methods was not satisfying since it didn't take the important temporal information into consideration and there may be many irrelevant frames confused the results. So, many new works are proposed to solve this problem, like Two-Stream[6], Long Short Term Memory(LSTM)[1] and 3D Convolutional Networks(3D-ConvNets)[7]. Two-Stream method uses frame images and multi-frame optical flow as two streams to train Convolutional Networks(ConvNets). LSTM uses memory cells to store, modifies and accesses internal states, so it is easier to find out the long-range temporal relationships between frames. Compared with ConvNets, 3D-ConvNets uses 3D convolutional kernels instead of 2D convolutional kernels. Thus the temporal information will remain in the output of such type of network.

1.2 Project Goals

The goals of this thesis project including:

1. Do interaction recognition based on the dataset UT-interaction. The input is the segmented videos and the output is the class labels for each test videos.

2. Do interaction detection based on the dataset UT-interaction. The input is the un-segmented videos and the output is the spatial and temporal location of specified classes.

3. Resemble one of the state-of-art methods, like 3D-ConvNet or CNN plus LSTM as baseline.
4. Construct an multi-stage network which use the first stage network to detect the body part pose/action and use the second stage network to detect the higher level interaction.
5. Find out a efficient training method to implement high performance networks with limited target training data-sets. This is refer to use pre-training technique to pre-train the networks with more general datasets.
6. Combine different networks to achieve higher performance.

1.3 Contributions

I will introduce the overall results and contributions of my thesis project.

CHAPTER 2

RELATED WORK

Briefly introduce the overview of related works. Since the main difference between those methodologies is video representation. So, I will introduce the main two technical categories: hand-crafted feature descriptors and neural-network based feature descriptors.

2.1 Hand-crafted Feature Descriptor

Begins a section.

2.1.1 SIFT-3D

Begins a subsection.

2.1.2 HOG-3D

Begins a subsection.

2.1.3 Improved Dense Trajectories

Begins a subsection.

2.2 Neural Network Based Feature Descriptor

Begins a section.

2.2.1 Single Frame 2D ConvNet

Begins a subsection.

2.2.2 LSTM

Begins a subsection.

2.2.3 3D ConvNet

Begins a subsection.

2.2.4 Two Streams ConvNet

Begins a subsection.

2.3 Datasets

Begins a section.

2.3.1 UCF101

Begins a subsection

2.3.2 Sports-1M

Begins a subsection

2.3.3 ActivityNet

Begins a subsection

2.3.4 CMU Panoptic Studio

Begins a subsection

CHAPTER 3

Architecture

Chapter 4

Design

CHAPTER 5

Experimental Results

CHAPTER 6

CONCLUSION

APPENDIX A

Begins an appendix

BIBLIOGRAPHY

[1] J. DONAHUE, L. A. HENDRICKS, AND S. G. ET AL., *Long-term recurrent convolutional networks for visual recognition and description*, CVPR, (2015).

[2] G. FLITTON, T. BRECKON, AND N. MEGHERBI BOUALLAGU, *Object recognition using 3d sift in complex ct volumes*, Proceedings of the British Machine Vision Conference, (2010), pp. 11.1–11.12.

[3] A. KARPATHY, G. TODERICI, S. SHETTY, T. LEUNG, R. SUKTHANKAR, AND L. FEI-FEI, *Large-scale video classification with convolutional neural networks*, CVPR, (2014).

[4] A. KLASER, M. MARSZALEK, AND C. SCHMID, *A spatio-temporal descriptor based on 3d-gradients*, British Machine Vision Association, (2008), pp. 275:1–10.

[5] P. SCOVANNER, S. ALI, AND M. SHAH, *A 3-dimensional sift descriptor and its application to action recognition*, 15th ACM, (2007), pp. 357–360.

[6] K. SIMONYAN AND A. ZISSERMAN, *Two-stream convolutional networks for action recognition in videos*, CVPR, (2014).

[7] D. TRAN, L. BOURDEV, R. FERGUS, L. TORRESANI, AND M. PALURI, *Learning spatiotemporal features with 3d convolutional networks*, ICCV, (2015).

[8] H. WANG AND C. SCHMID, *Action recognition with improved trajectories*, IEEE International Conference on Computer Vision, (2013).