VIDEO ANNOTATION USING CONVOLUTION NEURAL NETWORK

WAN ZAHIRUDDIN BIN W ABD KADIR

UNIVERSITI TEKNOLOGI MALAYSIA

VIDEO ANNOTATION USING CONVOLUTION NEURAL NETWORK

WAN ZAHIRUDDIN BIN W ABD KADIR

A thesis submitted in fulfilment of the
requirements for the award of the Master of Engineering (Electrical- Computer and Microelectronic System)”

Faculty of Electrical
Universiti Teknologi Malaysia

JANUARY 2018

Dedicated to my supervisor and friends. Specially dedicated to *Maa* and *Abah*

I really miss both of you.

Al-Fatihah

ACKNOWLEDGEMENT

During the accomplishment process on this project report, I had gained a lot of experiences and knowledge in the field of Computer and Microelectronics System. I owed these to a great deal of individuals. Therefore, I would like to use this opportunity to acknowledge and express my heartfelt gratitude to them.

My sincerest appreciation goes to my supervisor, Dr Usman Ullah Sheikh. His supervision, motivation and endless patience during the duration of this project had helped me to complete the requirements of this project.

Moreover, a great deal of appreciation goes to my fellow postgraduate friends, who always guided me to complete this work. I am indebted to many of my friends who had assisted me in my project. Their assistance, encouragement and contribution had enlightened me whenever I faced with any difficulties in my project.

Last but not least, my warmest regards to my family for their support and caring which allowed me to complete this journey.

ABSTRACT

In this project, the problem addressed is human activity recognition (HAR) from video sequence. The focussing in this project is to annotate objects and actions in video using Convolutional Neural Network (CNN) and map their temporal relationship using full connected layer and softmax layer. The contribution is a deep learning fusion framework that more effectively exploits spatial features from CNN model (Inception v3 model) and combined with fully connected layer and softmax layer for classifying the action in dataset. Dataset used was UCF11 with 11 classes of human action. This project also extensively evaluate their strength and weakness compared previous project. By combining both the set of features between Inception v3 model with fully connected layer and softmax layer can classify actions from UCF11 dataset effectively upto 100% for certain human actions. The lowest accuracy is 27% by using this method, because the background and motion is similar with other actions. The evaluation results demonstrate that this method can be used to classify action in video annotation.

ABSTRAK

Dalam projek ini, masalah yang ditangani adakah pengecaman aktiviti manusia (HAR) dari urutan video. Fokus dalam projek ini mengkelaskan tindakan dalam video menggunakan *Rangkaian Neural Convolutional (CNN)* dan memetakan hubungan antara gambar dengan mengunakan lapisan bersambung sepenuhnya dan lapisan softmax. Sumbangan ke atas projek ini adalah rangka kerja pembelajaran yang lebih berkesan untuk mengeksploitasi ciri-ciri spatial dari model CNN (model Inception v3) dan menggabungkan lapisan disambungkan sepenuhnya dan lapisan softmax untuk mengelaskan tindakan dalam dataset. Dataset yang digunakan untuk projek ini adalah UCF11 dengan 11 kelas aktiviti manusia. Projek ini juga menilai kekuatan dan kelemahan berbanding projek sebelumnya. Penemuan pada projek ini, yang menggabungkan kedua-dua set ciri antara Inception model v3 dengan lapisan bersambung dan lapisan sotfmax dapat mengkelaskan aktiviti dari dataset UCF11 secara berkesan dan mencapai 100% untuk aktiviti tertentu. Ketepatan terendah adalah 27% dengan menggunakan kaedah ini, kerana latar belakang dan gerakan serupa dengan tindakan lain. Hasil penilaian menunjukkan bahawa kaedah ini boleh digunakan untuk mengkelaskan tindakan dalam video.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

ANN	-	Artificial Neural Network
CNN	-	Convolutional Neural Network
HAR	-	Human Activity/Action Recognition
LSTM	-	Long Short Term Memory
PC	-	Personal Computer
PCA	-	Principle Component Analysis
RNN	-	Recurrent Neural Network
ReLU	-	Rectifier Linear Unit
SVM	-	Support Vector Machine
T-SNE	-	t-Distribued Stochastic neigbor Embedding
VOT	-	Visual Object Tracking
V3	-	Version 3

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Nowadays, the source of video comes from multiply sources for example, digital video cameras, internet (YouTube and Netflix), CCTV video, television, and others. Video processing is important for security, searching, education, movie and others. In current society where technology has already being developed rapidly, we cannot walk or drive around without being captured by security device or surveillance video. The data from video is important for user to classify the object and action in the video from CCTV, camera, video recording and other source. Besides that, video annotation is important for education especially for baby education, the video can show the object and action in video. For example Figure 1.1 shown the person, dog and chair, so the object can conclude the action of video with person playing with dog . This is easier for children to know the object in video.

Lately, many research on deep learning in terms of image and video processing. Deep learning is a new area of machine learning research, where deep learning is about learning multiple levels of representation and abstraction that helps to make sense of data such as images, sound, and text. Recently, deep learning is applied to many signal processing areas such as image, video, audio, speech, and text and has produced surprisingly good result. Convolutional Neural Network (CNN) deep learning is one of algorithm or method of the deep learning.

Video annotation is one of the most important application in video processing which is to annotate important objects and actions in video. Thus, video annotation using CNN deep learning approach is valuable for community right now. Automated object annotation in video is a crucial part within these applications and has become an important research area in image and video processing to ensure the minimization of using human control and having a faster response based on annotation in video. Figure 1.1, how the object will be annotate with box the object in the video.

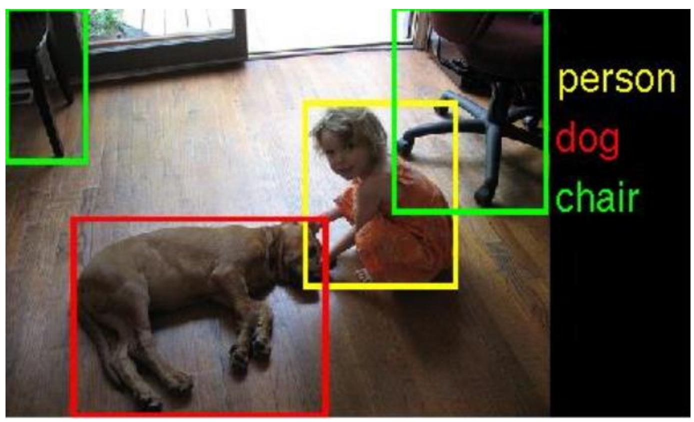

Figure 1.1: Person, dog and chair detected from a video

1.2 Problem Statement

Convolutional Neural Networks (CNN) have been developed and used in many areas such as security and surveillance to classify image or video content. CNN have been extensively applied for image and video processing problem such as recognition, detection, segmentation and retrieval. However, there are still many limitations of CNN for image and video annotation.

Among the challenges in video annotation using CNN algorithm is variable length on video processing. Normally, video segment are split into fixed chunks length which may span segment boundary. But variable length of video processing, the segment of video is also different and this is a challenging issue for CNN to handle it.

Besides that, temporal dependence of data is another problem in video annotation. The ability to manipulate the temporal dependencies in video data is important for a number of compressed domain video processing tasks. The difficulty on temporal dependencies of data is to develop a method for performing frame conversion on video processing. So that, this frame conversion are used to develop compressed domain video processing algorithm for performing temporal mode conversion [7].

Besides that, in video annotation the accuracy is most important because the object and action desired to annotate in video. In video, there are a lot of object, so to classify the desired object is difficult to determine, this is quite challenging for the video processing algorithm. Lastly, the large number of frames and high computational complexity is also one of the challenges in video annotation. This is because the large number of frames is time consuming.

Due to these challenging factors, the video processing flow must be considered before integrating with CNN to ensure the object in video can be annotated correctly. Thus, video annotation using CNN is a bit more complicated compare to image annotation using CNN.

1.3 Research Objectives

This research is about to annotating action based on object in video using convolutional neural network (CNN).. The goals of this project is as follows:

i. To annotate object for classifying action in video using CNN.
ii. To improve CNN implementation in terms of accuracy and performance when annotating the action on video.

1.4 Project Scope and Limitations

This project is focused on annotating object and classifying the action on video using convolution neural network (CNN) deep learning algorithm. The UCF11 dataset is used in this project to compare with previous research. This tools used are Tensorflow and python programming.

1.5 Thesis Layout

Chapter 2 reviews the literatures and previous works related to object and action detection in video. Chapter 3 focuses on the project design methodology which covers overview of the project flow and the algorithm flowchart. Chapter 4 presents the results and analysis of the works done throughout this research. Chapter 5 summarizes this research and gives recommendation for future wants.

REFERENCES

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Annual Conference on Neural Information Processing Systems*, 2012, pp. 1106–1114
2. C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Annual Conference on Neural Information Processing Systems*, 2013, pp. 2553–2561.
3. R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587
4. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks", In *Proc. CVPR*, pages 1725–1732, Columbus, Ohio, USA, 2014. 1, 2, 6, 7, 8
5. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2014). Translating Videos to Natural Language Using Deep Recurrent Neural Networks.
6. Vondrick, C., & Ramanan, D. (2011). Video Annotation and Tracking with Active Learning The era of big data, 1–9.
7. J. Y. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. "Beyond short snippets: Deep networks for video classification", *Conference on Computer Vision and Pattern Recognition*, 2015. 3, 4
8. V. N. Murthy, S. Maji, R.Manmantha, "Automatic Image Annotation using Deep Learning Representations". 2015
9. Mccann, S., & Reesman, J. (n.d.). Object Detection using Convolutional Neural Networks, 1–5.
10. jia, yangqing (no date) *Caffe*. Available at: http://caffe.berkeleyvision.org/
11. Girshick, R. (2015). Fast r-cnn. In: *International Conference on Computer Vision*, 1440–1448.
12. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast

feature embedding. In *Proc. of the ACM International Conf. on Multimedia*, 2014.pp 2

13. S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories”. In *Conference on Computer Vision and Pattern Recognition*, 2006. 1
14. Kiros, R., & Szepesvári, C. (2012). Deep Representations and Codes for Image Auto-Annotation. Advances in Neural Information Processing Systems 25, 1–9.
15. Wang, L., & Sng, D. (2015). Deep Learning Algorithms with Applications to Video Analytics for A Smart City: A Survey, 1–8.
16. D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, “Scalable object detection using deep neural networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2155–2162.
17. T.-Y. Lin, A. Roy Chowdhury, and S. Maji, “Bilinear CNN models for fine-grained visual recognition”, 2015.
18. A. Stepien, September 9, 2015, Deep Learning. Available at: http://semantive.com/deep-learning/
19. L. Deng, D. Yu, “Deep Learning Method and Applications”. Foudations and Trend in Signal Processing. 7:3-4
20. Girshick, R., Donahue, J., Member, S., Darrell, T., & Malik, J. (2016). ‘Region-based convolutional networks for accurate object detection and segmentation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*,2016, 38(1), 142–158.
21. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” 2014. [50]
22. C. Szegedy, S. Reed, D. Erhan, and D. Anguelov, “Scalable, highquality object detection,” arXiv e-prints, vol. arXiv: 1412.1441v2 [cs.CV], 2015.
23. H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In IEEE *Conference on Computer Vision and Pattern Recognition* , 2011, pages 3169-3176.
24. Laptev, I., Marszałek, M., Schmid, C., & Rozenfeld, B.. Learning realistic human actions from movies. 26th IEEE Conference on Computer Vision and Pattern Recognition,CVPR , 2008

25. S. Hochreiterand J. Schmidhuber, “Long short-term memory”. *Neural Computing*, 9(8):1735–1780, Nov. 1997. 2
26. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision.
27. Gammulle, H., Denman, S., Sridharan, S., & Fookes, C. (2017). Two stream LSTM: A deep fusion framework for human action recognition. In *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*(pp. 177–186). Institute of Electrical and Electronics Engineers Inc.
28. Zhang, Y., Liang, P., & Wainwright, M. J. (2016). Convexified Convolutional Neural Networks. *498 IEEE Transactions on Human-Machine Systems, Vol. 46, No. 4, 46*(4), 498–509. http://doi.org/10.1145/2951024
29. Sainath, T., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2015–August*, 4580–4584.
30. Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos, 1–9.
31. Hochreiter, S., & Urgen Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8),pp. 1735–1780.
32. Srivastava, N., Mansimov, E., & Salakhutdinov, R. (2015). Unsupervised Learning of Video Representations using LSTMs. Icml, 37,pp 2009.
33. “UCF11,” http://crcv.ucf.edu/data/UCF_YouTube_Action.php.