Proc. 1st International Conference on
Machine Learning and Data Engineering (iCMLDE2017)
20-22 Nov 2017, Sydney, Australia
ISBN: 978-0-6480147-3-7

A Deep Learning based approach for Human Action Recognition

Minhaz Uddin Ahmed[1], Kim Jin Woo[1], Miyoung Nam[1], Md Rezaul Bashar[2] and Phill Kyu Rhee[1]

[1] Department of Computer Engineering, Inha University, Incheon, South Korea
[2]Science, Technology and Management Crest, Sydney, Australia

Corresponding author's E-mail:pkrhee@inha.ac.kr

Abstract

Action recognition is gaining importance due to its numerous uses in security assistance, video analysis, and surveillance application. Action recognition from an image and online video is challenging works due to the complex background, object scale, variation of pose and miss interpretation of action. In this paper, we proposed a Deep learning based method for human action recognition using Convolutional Neural Network (CNN).The CNN use jointly learning feature transformation which classifies the object in image optimally and provides better performance on the other hand traditional manually designed features cannot perform well. Deep learning based approach exploits GPU as well as computational resources and produces a competent performance on real time videos and intelligent technology Lab dataset of Inha University. In our experiment, we use different dataset images for training such as simple and noisy and cluttered environment. We have trained the intelligent technology Lab dataset by pre-trained network (VGG16 network by the University of Oxford) model. We have found cropped ROI with higher iteration has better action recognition than large image while performing test owing to the less noisy background. Total five activities were considered due to the limited number of labeled training examples. We also integrated an active semi-supervised learning (ASSL) method which helps in improving the overall human action recognition and achieved average accuracy of 96%. For testing, we recorded videos which contain phoning, taking photo, using computer, jumping and reading activities. Our approaches significantly achieved higher MAP on intelligent technology Lab dataset.

Keywords: Action Recognition, Convolutional Neural Network, Deep Learning, Transfer Learning.

1. INTRODUCTION

Human action recognition is an important problem because through this we can identify the real life person interaction, person movement and particular task carry out by a person. Still picture or video contains significant information about human actions. Interpretation and understanding of these actions from images helps to resolve real life problems such as surveillance video analysis for safety, elder patient monitoring, sports video analysis, detection of abnormal activities or irregular behavior, human computer interaction and so on. All these applications require efficient action recognition from video. Due to complex environment such as crowded place (e.g market, bus station), pose dissimilarity (e.g running and

walking), person to person interaction (e.g playing), person to object interaction(e.g using phone or camera), low image quality , occlusion, noisy background, poor lighting condition and many other issues in the real time make it difficult to recognize human actions.

In order to solve this problem, we propose a Visual Geometry Group of Oxford University (VGG) network based deep learning model to recognize the human action with integrated Active semi-supervised Learning (ASSL) framework.
In recent years, Deep Convolutional Neural Network has been extensively used for computer vision applications especially human action detection by C. Feichtenhofer, A. Pinz, A. Zisserman (2016), G. Gkioxari, R. Girshick, and J. Malik (2015), H. Mahmudul, and Amit K. Roy-Chowdhury (2015) because of higher accuracy and appropriate application to real life problems. Over the last few years object recognition method become more developed than previous approaches such as the Deformable Part Model(DPM) by P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan (2010) , AlexNet by A. Krizhevsky, I. Sutskever, and G. Hinton (2012), VGGNet by Karen Simonyan, Andrea Vedaldi ,Andrew Zisserman (2013). We deploy VGGNet in our research work due to its promising results. Some other popular methods are based on a probabilistic model by Lei Zhang, Zhi Zeng, and Qiang Ji (2011) which works well among simple actions in a favorable environment with HMM by Hanju Li ,Yang Yi , Xiaoxing Li , Zixin Guo (2013) method. However, very complex action in a dynamic environment often impossible to detect correct action such as brushing hair and using the cell phone or fight and hug in an occluded environment.

In the next section we analysis the literature of action recognition in still images and videos. We describe our approach in section 3 action recognition based on transfer learning. In section 4 we describe experiments carry out based on the Intelligent Technology Lab action dataset. In section 5 we explained our experiment result and contribution.

2. RELATED WORK

Action classifications, activity recognition, human behavior analysis in image sequence have been widely studied in computer vision area for last two decades. Considerable advancement also made in person detection in image such as space time interest point (STIP) by I. Laptev (2005), Histogram oriented gradient (HOG) by N. Dalal and B. Triggs (2005), Bag of words (BOW) by V. Delaitre, I. Laptev, and J. Sivic (2010), Dense trajectory based approach by Heng Wang, Alexander Klaser, Cordelia Schmid, Cheng-Lin Liu (2011), HMM based approach. Most recently CNN based approach by C. Feichtenhofer, A. Pinz, A. Zisserman(2016), G. Gkioxari, R. Girshick, and J. Malik (2015), H. Mahmudul, and Amit K. Roy-Chowdhury (2015), Karen Simonyan, Andrea Vedaldi ,Andrew Zisserman (2013) gained huge popularity due to the outstanding performance in PASCAL VOC and ImageNet challenge by A. Berg, J. Deng, and L. Fei-Fei (2010). These two benchmark dataset contributed a lot to improve the of object recognition in image and video file. Additionally, RCNN by G. Gkioxari, R. Girshick, and J. Malik (2015), Ross Girshick (2015), convolutional two-stream network by C. Feichtenhofer, A. Pinz, A. Zisserman(2016), continuous learning framework by H. Mahmudul, and Amit K. Roy-Chowdhury (2015) played important role while in our work we focus on semi-supervise learning with active learning approach.

Each image file with object bounding box information stored in separate xml file has gain popularity in last several years if we consider the object detection competitions such as PASCAL VOC, imageNet, MSCOCO. Bounding box keeps the region of interest information of particular object in an image. In our work we used Fast-RCNN by Ross Girshick (2015) to detect person.
Action classification specifies the relation between object and person orientation in a scene. In our work we focus on human with nearest object to find out the action type in a picture. There are numerous method have been used to recognize action. However, CNN showed promising result by G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik (2014). In our work, we adopted VGG model where transfer learning played the pivotal role in our experiment. By using transfer learning we can re-train pre-existing neural network (e.g. AlexNet, VGGNet) minor modification and then retrain it in our experiment for action classification by Sinno Jialin Pan and Qiang Yang (2010). It is an efficient way to apply deep learning to solve challenging problem.

3. METHOD

We consider two steps approach for action recognition. First of all, we detect person through Fast RCNN by Ross Girshick (2015) algorithm for each class and then later apply our ASSL framework for increasing human action recognition. Our ASSL method is a state of the art method work in two steps, first step is Semi Supervised Learning (SSL) then in second step we consider Active Learning (AL) method which ensures greater data justification with higher accuracy. Here we apply transfer learning approach where transfer knowledge from earlier learned the task.

Figure 1. Classified action in images with noisy background and shows recognition score at the top.

In this work we propose Active semi supervised Learning (ASSL) framework for action recognition from video and single image as shown in figure 1. We anticipated the action specific person detection where each image contains person information with bounding box data with action stored in xml file. We train the image dataset and make our initial model using openly available VGG16 model and modify the last two layer to adopt for human action recognition. If the initial model cannot predict the right action, we employ active learning method to increase the performance.

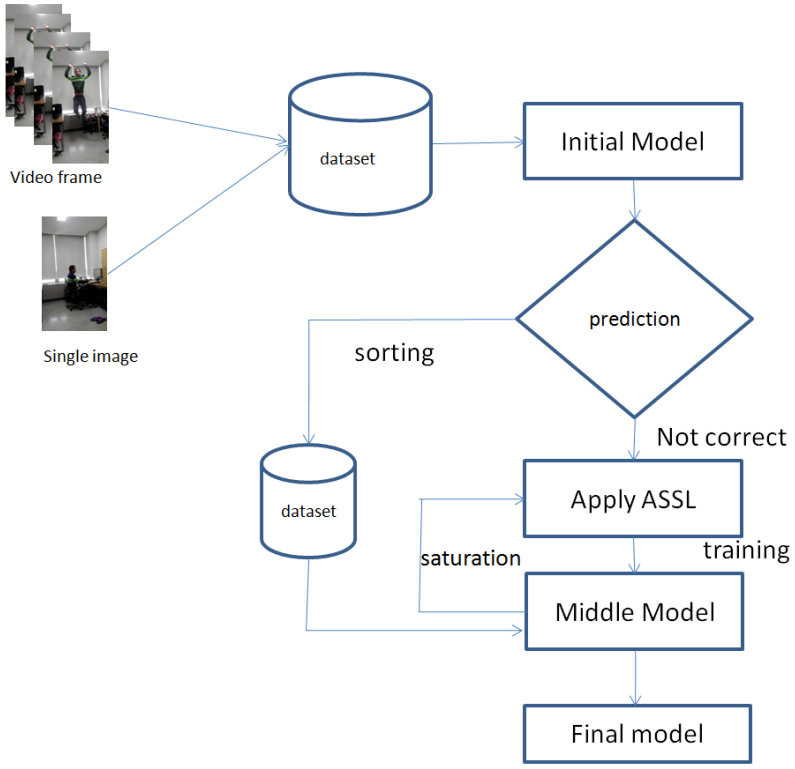

Figure 2. Entire system diagram

We prepare number of training and test image batches, we check the performance against the initial model. For the performance checking we consider threshold value 0.9. If maximum number of image of a batch has wrong prediction as well as less than threshold value we skip that batch and replace with new batch of image and train again. In similar way, we apply incremental learning and prepare the final model which classifies the correct human action.

4. DATASET

The datasets for human action recognition are different from each other due to dissimilar annotation structure. At the same time manual annotation with ground truth is time consuming and tedious work. The pre-trained network was trained on these datasets with image annotation. For action recognition work we use Intelligent Technology Lab dataset. To train the dataset which consist several thousand images where actions are labeled with bounding box information. There are total five different actions such as jumping, phoning, taking photo, using computer and reading. We use the intelligent technology Lab test dataset to evaluate for action classification. The action categorized work presumes knowledge of the ground truth location of the person throughout test time. For the action classification evaluation we use the evaluation criteria defined by the PASCAL VOC by Mark Everingham (2015) action task which computes the AP on the ground truth test boxes.

5. EXPERIMENTS

In this section, we describe the details about our experiments action recognition. We use the CNN architecture by A. Krizhevsky, I. Sutskever, and G. Hinton (2012) which worked extremely well in object classification on ImageNet challenge. We construct on VGG16 network Karen Simonyan, Andrea Vedaldi, Andrew Zisserman (2013) which showed a remarkable improvement from previous state of the art methods.

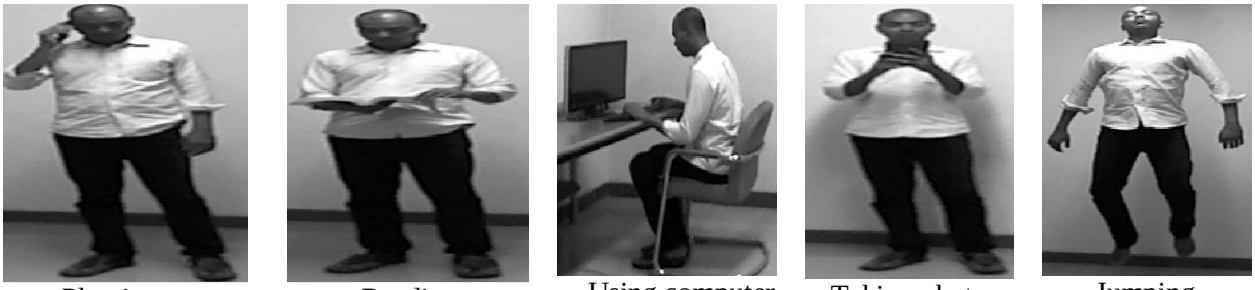

Figure 3. ITLab action dataset where background is simple.

Figure 3. Shows preprocessed action dataset such as phoning, reading, using computer, taking photo and jumping where environment is less noisy. Here pre-processing includes intensity normalization and ROI selection.

5.1 Implementation Details

Our experiment use publicly available VGG net by Karen Simonyan, Andrea Vedaldi and Andrew Zisserman (2013) (http://www.vlfeat.org/matconvnet/pretrained/) that has 5 convolutional layers and 3 fully connected layers. These networks are pre-trained ImageNet for 1000 categories classification task. For the person detection ASSL use the Fast-RCNN by Ross Girshick (2015) that is convolutional neural network based object detector and ASSL frameworks is implemented on the popular deep learning library Caffe. All implementations are on a single server with and CUDA 7.5 with cuDNN 5.1 and a single NVIDIA GeForce GTX 970. We also use Matlab 2015b with matconvnet by Andrea Vedaldi and Karel Lenc (2015) library for both windows 7 and Ubuntu 14.4 operation system in our experiment. We train our initial model for 20k iterations with batch size 8 and learning rate of 0.001. The average single batch of 25 image with 50k iteration batch size 8 using GPU takes nearly 15 minutes.

Table 1. Number of images ITLab Action dataset

	Total	**Average Accuracy**	**Total**	**Average Accuracy**
Training Image	502	94%	621	96%
Test Image	50		50	

Table 2. Number of detected action in the training and test sets in ITLab action dataset.

	Jumping	**Phoning**	**reading**	**Taking photo**	**Using computer**
Train	100	100	100	100	100
Test	10	10	10	10	10
Correctly recognized image	9	10	10	8	10

We evaluate our experimental result in ITLab dataset. Our propose system, depends on large amount of data for training the model. We compare our results against state of the art methods which show that action recognition method with active semi supervised learning yields superior performance on ITLab dataset.

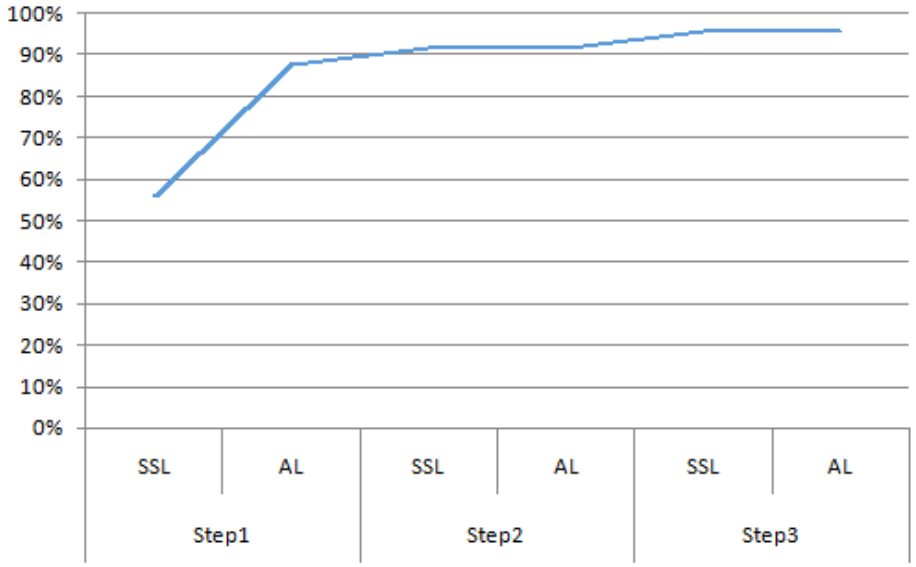

Figure 4. Shows the Semi Supervised Learning (SSL) and Active Learning (AL) steps for Action recognition.

In our experiment, each step consists of two sub-steps AL and SSL. In figure 4. Step one show Semi-Supervised Learning (SSL) get 56% accuracy while Active Learning (AL) obtains 88% correctness. In step two with combined dataset with higher iteration SSL and AL get 92 precision. Finally, in step three both SSL and AL achieve 96% accuracy.

6. CONCLUSION

In this paper we presented a deep learning based framework called ASSL (Active Semi-Supervised Learning) which improves the action recognition performance and reduce the burden of human label effort. Looking at the result, we see that Deep Convolutional Neural Network is suitable for human detection and activity recognition on different ITLab action datasets, but they still require much more tuning with large volume of dataset for better accuracy. In future we would like to compare with other state of the art benchmark dataset with higher performance.

ACKNOWLEDGEMENTS

This work was supported by the ICT R&D program of MSIP/IITP. [2017-0-00543], Development of Precise Positioning Technology for the Enhancement of Pedestrian's Position/Spatial Cognition and Sports Competition Analysis. The GPUs used in this research were generously donated by NVIDIA.

REFERENCES:

C. Feichtenhofer, A. Pinz, A. Zisserman (2016). Convolutional Two-Stream Network Fusion for Video Action Recognition

G. Gkioxari, R. Girshick, and J. Malik (2015). Contextual action recognition with R*CNN. In ICCV.

P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan (2010).Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645.

H. Mahmudul, and Amit K. Roy-Chowdhury (2015). A Continuous Learning Framework for Activity Recognition Using Deep Hybrid Feature Models. Multimedia, IEEE Transactions on 2015.

A. Krizhevsky, I. Sutskever, and G. Hinton (2012). ImageNet classification with deep convolutional neural networks. In NIPS.

Karen Simonyan, Andrea Vedaldi, Andrew Zisserman(2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps arXiv: 1312.6034,

Lei Zhang, Zhi Zeng, and Qiang Ji (2011). Probabilistic Image Modeling With an Extended Chain Graph for Human Activity Recognition and Image Segmentation, IEEE transactions on image processing, vol. 20, no. 9

Hanju Li, Yang Yi , Xiaoxing Li , Zixin Guo (2013). Human activity recognition based on HMM by improved PSO and event probability sequence, Journal of Systems Engineering and Electronics, Volume: 24, Issue: 3.

I. Laptev (2005). On Space-Time Interest Points, in International Journal of Computer Vision, vol 64, number 2/3, pp.107-123.

N. Dalal and B. Triggs (2005).Histograms of oriented gradients for human detection, in Proc. IEEE Comput. Soc. Conf. CVPR, Jun. 2005, pp. 886–893.

V. Delaitre, I. Laptev, and J. Sivic (2010). Recognizing human actions in still images: a study of bag-of-features and partbased representations. In BMVC 2010.

Heng Wang, Alexander Klaser, Cordelia Schmid, Cheng-Lin Liu (2011).Action Recognition by Dense Trajectories, Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference

Ross Girshick (2015). Fast R-CNN - ICCV 2015

G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik (2014). R-CNNs for pose estimation and action detection. arxiv.1406.5212

LeCun, Y., Bottou, L., Bengio, Y (1998) Gradient-based learning applied to document recognition, Proc. IEEE, 1998, 86, (11), pp. 2278–2324

A. Berg, J. Deng, and L. Fei-Fei (2010). Large scale visual recognition challenge (ILSVRC), 2010. URL http://www.image-net.org/challenges/LSVRC/2010/.

M. Ranjbar, T. Lan, Y. Wang, S.N. Robinovitch, Ze-Nian Li, and G. Mori (2013). Optimizing Nondecomposable Loss Functions in Structured Prediction, IEEE transactions on pattern analysis and machine intelligence, vol. 35

B. Yao and L. Fei-Fei (2010). Modeling mutual context of object and human pose in human-object interaction activities. In CVPR.

Gary Ge, Kiwon Yun ,Dimitris Samaras, Gregory J. Zelinsky (2015). Action Classification in Still Images Using Human Eye Movements, IEEE

Fahad Shahbaz Khan, Jiaolong Xu, Joost van de Weijer, Andrew D. Bagdanov, Rao Muhammad Anwer, and Antonio M. Lopez (2015).Recognizing Actions through Action-Specific Person Detection, IEEE transactions on image processing, vol. 24, no. 11.

S. Ren, K. He, R. Girshick, and J. Sun (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, arXiv:1506.01497.

Sinno Jialin Pan and Qiang Yang (2010). A Survey on Transfer Learning, IEEE transactions on knowledge and data engineering, vol. 22, no. 10.

Andrea Vedaldi, Karel Lenc (2015). MatConvNet: Convolutional Neural Networks for MATLAB, Proceedings of the 23rd ACM international conference on Multimedia

P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan (2010).Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman (2015). The PASCAL Visual Object Classes Challenge: A Retrospective, International Journal of Computer Vision