ACTION DETECTION IN OFFICE SCENE BASED ON DEEP CONVOLUTIONAL NEURAL NETWORKS

SHI-YANG YAN[1], YU-DI AN[1], JEREMY S. SMITH[2], BAI-LING ZHANG[1]

[1]Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China
[2]Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, L69 3BX, UK
E-MAIL: Shiyang.Yan@mail.xjtlu.edu.cn

Abstract:

In many scenarios, a persons behavior in office environment needs to be monitored and some predefined abnormal actions or activities should be detected and recognized. In this paper, we attempted towards the solution starting from a persons pose with poselets as the basic building blocks. The existed powerful pose representation, i.e., poselets, together with deep convolutional neural networks, are exploited to implement an efficient action recognition system from still images. The system extends poselets detector to region proposal, cascaded with R*CNN for final action detection. Unlike many published work which only emphasis on action classification, our system implements multi-task learning with classification and localization of person and the corresponding actions simultaneously, To facilitate our studies, a specially designed action dataset was created. Preliminary experiments demonstrates promising results.

Keywords:

Abnormal behavior alarming; Action detection; Poselets; Convolutional neural network

1 Introduction

Human behavior in office environments often need to be monitored with video cameras for security reasons. For example, in order to prevent commercial espionage and protect intellectual property, many companies do not allow employees or visitors taking photos in office. This is a special topic of action and activity recognition, which have been endeavored by scientists from computer vision for many years.

While most of the previous works on action recognition were centered around spatio-temporal patterns based on motion/video datasets, a number of recent studies have paid attentions on the detection and recognition of actions or activities based on still images [1]. Such explorations are worthwhile as still images are prevalent. However, there are more challenges as the detection of humans and their corresponding poses is more difficult than in video and the definition of some actions might become ambiguous if the motion direction is not available [2].

In this paper, we attempted to find solutions for the recognition of office behavior within the framework of deep convolutional neural networks and develop a novel approach to detect action in still image. Our emphasis is on the detection and recognition of human pose, i.e. the configuration of body parts, in the image, which is the most important cue for understanding human actions. As not all body parts are equally important for differentiating various actions [3], how to better discriminate actions by utilizing pose information is a complicated issue.

Our work aims at further exploration of poselet for office behavior analysis in the deep convolutional neural network framework. Poselets [4] was firstly proposed by Bourdev for the notion of parts, constructed to be tightly clustered in both appearance and configuration spaces. As an effective part-based method, poselets have been applied to person detection [5]. In the last several years, Convolutional Neural Networks[6] have been extensively studied in image recognition and other relevant tasks, often with state-of-the-art performance. Especially, the region CNN (R-CNN) proposed by Ross B. Girshick [7] has become a milestone of object detection and a number of subsequent variants have been proposed and applied in different vision problems. In [8], Ning Zhang applied poselets and deep convolutional neural networks for attribute discovery, with features firstly extracted from each type of poselets and then concatenated for the final attributes detection.

As our emphasis is on human actions in office environment which require hand movements during the time the actions are operated, we proposed a novel approach which features

person object proposal and corresponding poselets as inputs to R*CNN [9], a recently proposed improvement over R-CNN to use more than one region for classification while still maintaining the ability to localize the action. Unlike the approach in [8], we do not utilize every detected poselets for the feature representation. Instead, a more discriminative way is investigated for the selection of top contributors for a specified action in each of the poselets . Our work is mainly inspired by [9], in which contextual cues are exploited with Convolutional neural networks for action and attribute recognition. We treated the basic pose component, namely the poselets, as the contextual cues.

Our main contributions are as follows:

1. We proposed a novel system based on the deep convolutional neural network for action detection in office environment.

2. Instead of training poselets detector which is usually very demanding for the preparation of training data, we directly applied the open-source poselets detector thus largely reduced the workload.

3. Unlike original R*CNN which only classifies actions, we implemented human detection and corresponding action recognition simultaneously.

4. By introducing the combination of poselets and R*CNN, we extended the scope of contextual cues to basic pose component, which plays a critical role in action recognition.

2 System overview

The part-based model, i.e., poselets, and deep learning for acquiring discriminative representation of actions in still images, are at the core of our system for the recognition of human behavior in office environment. While poselets is used for the representation of human pose which is instrumental to the action cues, the potential of convolutional nets for representational learning is fully developed by selecting the most influential parts of human in the context of action recognition. As illustrated by Figure. 1, we combined poselets-based person detector and R*CNN to form a cascaded system, which accomplish action classification with promising result. Moreover, we delved into the multi-task learning strategy inherited from fast R-CNN [10], to achieve action classification and person localization at the same time.

2.1 Brief introduction of Poselets

In [4], Bourdev and Malik introduced a new notion of parts, poselets, constructed to be tightly clustered both in the configuration space of keypoints, as well as in the appearance space of image patches. The original poselets in [4] only use 3D configuration. Then, in [5], same authors extend it to 2D annotations and achieve person detection based on poselets. It is operated as sliding-window detectors on top of low-level gradient orientation features, such as HOG, and achieve recording results on person detection.

2.2 Brief introduction of R*CNN

Recently, with the excellent performance of deep convolutional neural networks in object detection and image classification, many works have attempted to apply CNN models in action recognition [11] [9]. Among them, R*CNN [9] is a representational work, which make use of all available contextual cues for final action recognition. These contextual cues, are either the object that interact with person or specific human parts that contribute mostly to the action. One of the novelties of R*CNN lies in the region proposal which is composed of two parts, namely, the primary region and secondary region. The primary region, is the ground truth, which provides person object in question. And the secondary region, which is all the bounding boxes generated by region proposal algorithm, provides contextual cues for action recognition. The region proposal is implemented by selective search [12]. Figure. 2 gives the basic structure of original R*CNN. Given an image I, we selected the primary region which contains person (red box) and the region defines set of candidate region proposals for contextual discovery. The inputs of R*CNN includes primary region that contains the person in question and secondary region that discover contextual cues. The process of secondary selection can be further illustrated as follows: R*CNN is implemented based on Fast R-CNN [10]. Our approach is different from the original R*CNN as we replace the selective search with poselets detector. Based on the multi-task strategy in fast R-CNN, we also improve the system in the way that our system is able to detect person and corresponding action simultaneously.

3 Implementation details

3.1 The action dataset acquisition

Though there exists a number of different human action or activities datasets available from the Internet, none of them was specially designed for office environment. To facilitate the study of recognizing complex human actions in a realistic surveillance setting and in an office environment, we created a dataset which focus on human-mobile interactions. Total 16 volunteers participated in establishing this dataset. All of the

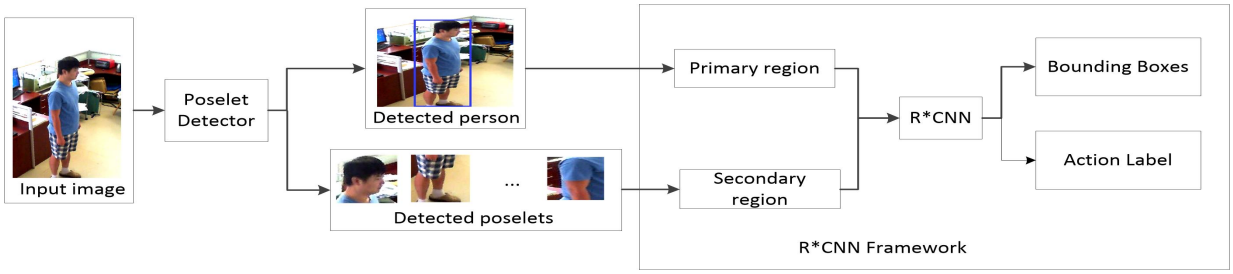

FIGURE 1. System structure: We treat detected person bounding boxes as primary region and poselets hits which support corresponding bounding boxes as secondary region, R*CNN with multi-task strategy is applied for final action detection and bounding box localization.

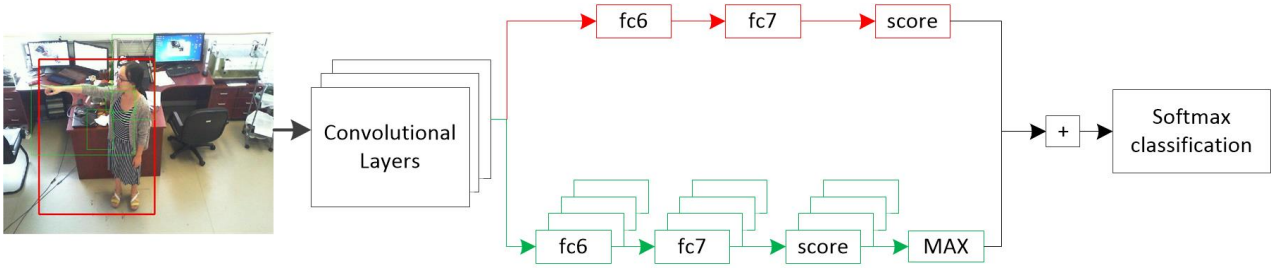

FIGURE 2. Principle of R*CNN: With primary region and secondary region, R*CNN process them using certain number of convolutional layers, then primary region will be classify subsequently while the top classification score of secondary region will be selected. After an operation of summing, a softmax classification is connected as final output.

subjects kept standing position with different poses while exercising a set of prescribed actions, including standing without doing anything, making a phone call, photographing and finger pointing. Figure. 3 gives some of the examples.

A number of factors have been taken into consideration in the data acquisition, including the background complexity, illumination and viewpoint changes. We exploited three cameras with different shooting angles. Figure. 3 further explains the setting for the data acquisition. As our focus is on action recognition from still images, we manually selected key frames from the recorded video. This will be replaced by an automatic algorithm in our next stage of works.

3.2 Implementation platforms

Our experiment was conducted on a dell Tower 5810 with Intel Xeon E5-1650 v3 and memory 64G. In order to speed up CNN training, a GPU, NVIDIA GTX TITAN, is plugged on the board. The program is operated on the 64-bit Open-source Linux operating system CentOS 7, CUDA 7.5, Python2.7.3 and Matlab 2014b linux version and newest version Caffe deep learning platform.

Specifically, following the common practice of pre-training CNN [13], the parameters of the layers of CNN-M from [14] is assigned. Then the CNN model is fine-tuned with region proposal from poselets detector. During training, we provide ground truth label of action class as well as region of a person. The max iteration of training in Caffe is set as 40000, which takes about 4 hours in our platform.

3.3 Training data preparation

The training was mainly facilitated by a ground truth dataset labelled for the defined action classes and the bounding box for a person. The original work on poselets[4] applied SVMs with HoG descriptors as poselets classifiers. Poselets training needs annotations of key points of human (eyes, ears, joints, nose) which is very expensive in terms of human resources and time consumed. As a result, inspired by bootstrapping method in [11], we applied an open-source poselet detector directly on our datasets. First of all, we annotate target person in each image only with person locations. Hence, we obtained a dataset with ground truth annotations. More specifically, each object hypothesis for person generated by the poselets is matched against the bounding box surrounding a person from the ground-truth, and hypotheses that overlap by more than 0.3 intersection-over-union (IoU ratio) are selected as training samples for the deep convolutional neural networks. All activated poselets hits that

FIGURE 3. Examples of our dataset

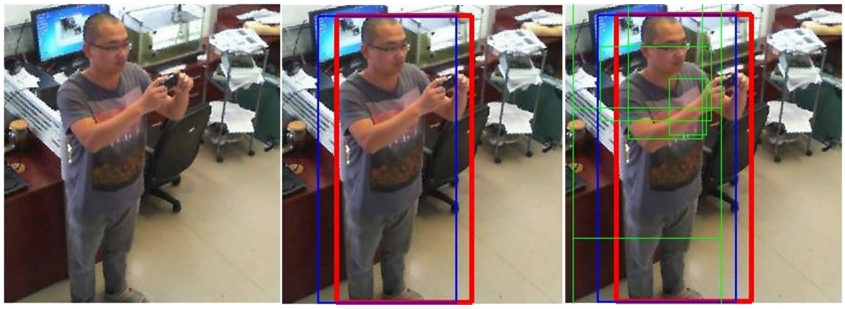

FIGURE 4. Illustration of data preparation: The left image is the original image; The middle image show the groundtruth(Red box) and detected bounding box(Blue box) which overlaps more than 0.3 with groundtruth region, and this is for our primary region during training; The right image gives examples of poselets hits(Green boxes) that support the selected bounding box.

support the above-mentioned bounding boxes are treated as secondary region proposal in R*CNN. Figure. 4 illustrates the process of training data preparation. This bootstrapping mechanism allowed us, without significant annotation effort, to collect all the training samples for R*CNN, largely improve the efficiency.

This bootstrapping mechanism allowed us, without significant annotation effort, to collect all the training samples for R*CNN, largely improve efficiency.

4 Experimental result

4.1 Evaluation of poselets detector performance

To evaluate the performance of open-source poselets person detector, we followed the basic approach in [15]. Specifically, we plotted the recall versus intersection-over-union (IoU) curve based on different detection threshold.

As shown in Figure. 5, we obtained recall value when the IoU threshold is 0.5. The best recall is when we extracted top 20 percent of detected person bounding boxes. Further decreasing threshold does increase recall performance, although it is marginal. Also, increasing bounding boxes for primary region and poselets for secondary region will increase the workload for final action classification and bounding box localization in CNN. Top 20 percent bounding boxes provide about 15 bounding boxes on average per image and 150 corresponding poselets hits per image. Consequently, in this paper, we extracts top 20 percent bounding boxes as the primary region and poselets hits that support the bounding boxes as the secondary region of R*CNN.

TABLE 1. Final result of action classification

AP(%)	standing	phoning	photographing	pointing	mAP
Our method	96.17	97.32	98.00	96.00	96.87

TABLE 2. Final result of action detection

AP(%)	standing	phoning	photographing	pointing	mAP
Our method	88.63	90.90	92.12	89.32	90.24

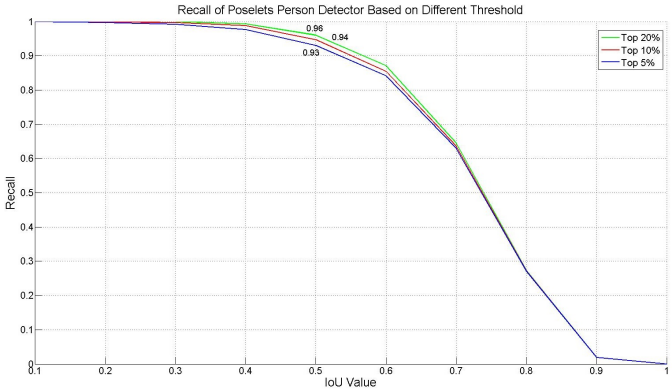

FIGURE 5. Recall of poselets person detector

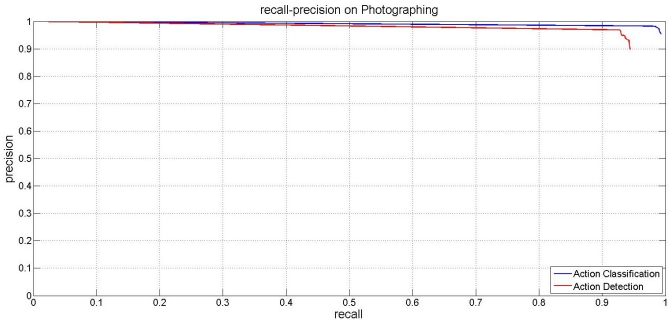

FIGURE 6. Recall-precision curve on photographing

4.2 Evaluation of system performance

There are two parts in system performance evaluation: Action classification and Action detection. We will evaluate the two tasks separately. For the task of action classification, the location of the people performing action is considered known. During test time, a ground truth region is being marked with a score for each action. We will evaluate the classification performance by calculating the Average Precision(AP) for each class. TABLE. 1 gives AP result on action classification of each class. The mAP of 96.78 is finally achieved.

The task of action classification assumes knowledge of the location of the people. This makes the task a lot easier, since it skips the difficult step of detection. A real application can not assume perfect localization of the objects to be classified. On the other hand, action detection is the task of localizing and classifying the people who is performing one of the actions. A correct prediction is the one that overlaps more than 0.5 with a ground truth instance and predicts the correct action label. Also, the AP value of each class and mAP will be calculated accordingly. This task is much more difficult as in some occasions even the correct action label is obtained the corresponding bounding box might not match the person. In this situation, the samples cannot be treated as true positive. TABLE. 2 shows AP result on action detection of each class. The mAP of 90.24 is finally achieved, which only have limited decrease compared with action classification. This result, consequently, imply our bounding box regression is powerful.

For more straightforward illustration, we also present the recall-precision curve of the class "photographing" in Figure. 6. As can be seen from the Figure. 6, both classification and detection achieve promising result.

5 Conclusion

In this paper, we propose to combine poselets person detector and R*CNN for action classification and detection on a dataset we created to simulate some common actions in office environment, and the interactions with mobiles in particular. By introducing a bootstrapping method for the preparation of training data, considerable time has been compromised for an action recognition system with convolutional neural networks work. Furthermore, we improve the poselets-dependent features by applying R*CNN to select top contributors of activated poselets. By extending the meaning of contextual cures from original R*CNN to some fundamental parts of a pose, we reached a more efficient system. Based on these improvements, promising results have been provided on both of action classification

and detection tasks. Future works include the experiments of our method on pubic datasets.

References

[1] G. Guo, A. Lai, A survey on still image based human action recognition, Pattern Recognition 47 (10) (2014) 3343–3361.

[2] M. R. Ronchi, P. Perona, Describing common human visual actions in images, arXiv preprint arXiv:1506.02203.

[3] W. Yang, Y. Wang, G. Mori, Recognizing human actions from still images with latent poses, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 2030–2037.

[4] L. Bourdev, J. Malik, Poselets: Body part detectors trained using 3d human pose annotations, in: International Conference on Computer Vision (ICCV), 2009.

[5] L. Bourdev, S. Maji, T. Brox, J. Malik, Detecting people using mutually consistent poselet activations, in: European Conference on Computer Vision (ECCV), 2010.

[6] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012, pp. 1097–1105.

[7] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[8] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev, Panda: Pose aligned networks for deep attribute modeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1637–1644.

[9] G. Gkioxari, R. Girshick, J. Malik, Contextual action recognition with r* cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1080–1088.

[10] R. Girshick, Fast R-CNN, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015.

[11] G. Gkioxari, R. Girshick, J. Malik, Actions and attributes from wholes and parts.

[12] K. E. Van de Sande, J. R. Uijlings, T. Gevers, A. W. Smeulders, Segmentation as selective search for object recognition, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 1879–1886.

[13] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[14] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, arXiv preprint arXiv:1405.3531.

[15] J. Hosang, R. Benenson, P. Dollr, B. Schiele, What makes for effective detection proposals?, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (4) (2016) 814–830. doi:10.1109/TPAMI.2015.2465908.