Commentary: <Recognizing Actions Through Action-Specific Person Detection>

<Guanzhu Hou>
<z5325417>

1. Introduction

Recently, human-computer interaction and abnormal behavior analysis through action recognition has become a hot research. Some popular practical examples, such as control of home appliances through gestures [1], the establishment of optimal sports model for athletes [2] and description of the user's image on social media to push advertising has fully proved the business prospect and broad application of action recognition. This paper focused on applying the recognition in static pictures.

The limitation for the current two-step methodology ignores that human actions deeply condition the poses in which humans are expected to be observed. And the lack of the number of examples in each category in the database also increases the difficulty of learning an accurate human model. Here the authors tend to solve the problem by using action-specific person detectors and posing this detection as a transfer learning based on the generic person detection by Deformable Part Model (DPM).

If the claim in this paper is valid, the full utilization of action class information would lead to more accurate person localization. After that, transfer learning will bring more robustness and propose higher quality bounding boxes to subsequent action classification on convolutional neural networks (CNNs). What's more, I think a still picture with accurate action recognized could an inspiration to video-based action recognition. For example, [3] is to sum the color heat maps on all static frames to obtain the potion representation of the whole video clip.

2. Methods

The authors divide the person proposals for action recognition into 3 parts: *direct learning of action-specific person detectors, transfer learning of action-specific person detectors and action classification with CNNs*.

Starting with the direct learning part, version 5.0 of DPM in [4] is performed in person detection, where the DPM parameters w^a are learned from latent SSVM optimization. a shows the link with human actions, that for each action type a, the model w^a will use labeled person instances executing action a as positive examples independently.

Although it does not give a detailed description of the DPM model it uses, with only mentions that it is a version of a doctoral thesis. I think the acceleration method of cascade DPM in [5] is also worthy of reference. As the calculation of DPM parameters will take more than 10 seconds each picture in Pascal VOC, this quadrupling the speed can avoid the bottleneck in practical applications. In addition to the improvement of speed, [6]'s method of introducing weakly supervised learning and roughly classifying candidate windows into target objects or non-target classes also significantly optimizes the DPM model. As summarized in [7], the state-of-the-art (SOTA) of the three series of pedestrian detectors (DPM, decision forests, deep networks) is unexpectedly close. Deepen the understanding of what makes good features better, to design better features, which will be more meaningful in future work. For example, [8] combines the image to image translation model in Generative Adversarial Network to generate the corresponding optical flow diagram for the still image, and captures the motion information that is generally considered to be reflected only by video. This enables the rich dynamic structures and actions contained in the image to be better utilized in practice than DPM of this paper, for example, to depict and infer the user's image on social media. Of course, as the conclusion in [7] has been made for almost 10 years, it is worth trying to replace the combination of DPM and latent SVMs with faster-RCNN or Yolo. Both two of the deep networks will also include the target with the bounding box, and has fast detection speed.

Moving onto the *transfer learning of action-specific person detectors*, the DPM parameters w^G learned from the generic person detector is about to use under the guidance of transfer learning. The idea of transfer learning comes from domain adaptation to eliminate the difference in probability distribution between test and training sets caused by changes in sensors or application environment in the acquisition process. In this paper, one of the inputs of the Adaptive SSVM (A-SSVM) is w^G and the other is the small subset of training examples. It is claimed that the result new w^a would be a good solution to limited training examples. However, it is not perfect, as the recall curve graph shows, in the category "running", transfer learning is not even as good as direct learning. I think data augmentation should be applied instead of only trust transfer learning. With the application of noise, inversion, scale change, and other means, the data scale can be controlled within an acceptable range, and the model can get more accurate results in the category "running".

Lastly, we discuss *classification with CNNs*. After getting results from a specific action-specific detector of each action type, the authors perform the classifier trained for the same action type. Here the 16-layer CNNs model (also test in 19-layer) applies 3×3 receptive fields with stride equals to 1.

For the strength, I think small convolution will save t the computational cost, which is similar to the shifting of the 7×7 convolution to 3 3×3 convolutions in ResNet-C [9]. For the weakness, with the continuous development of deep learning, this CNN model has too many alternatives, such as DenseNet, SENet or Optimized ResNet in [9]. It is necessary to judge what model to use through experiments, as SOTA model is not necessarily the most suitable.

3. Results

The authors divide the experiment results into 3 parts: *datasets and evaluation, action detection, and action classification*.

Beginning with *datasets and evaluation*. Both Stanford-40 and PASCAL VOC 2012 have been tested in this paper. Unlike deep learning, where 80% of the data set is usually used as the training set, nearly half of Stanford-40 are used for training. As the largest known still picture action recognition library, Sanford 40 has a relatively simple background. The PPMI data set also from Stanford has a more complex background and occlusion intuitively. In my opinion, this 4092 image library could also be selected to test whether the model can effectively deal with complex background and occlusion. Or we could use Soft Margin Support Vector Machines to handle noisy data instead of the hard margin SVM in this paper. For the PASCAL 2012, additional annotations were applied to increase the available data for training. Continuing with the evaluation, here in this paper, the authors use the average precision (AP) following the standard PASCAL detection protocol. I prefer the F1 score, for not only precision but also recall is used.

Moving onto *action detection*, on the Standford-40 dataset, the transfer learning method in this paper has a mean AP for 45.4%, and 31.4% on PASCAL VOC 2012. Compared with the DPM model trained for each action type (before Transfer-Learning, named Direct-Specific), the final Transfer-Learning model improves 7.8% and 2.8% respectively. For the per-category performance, the final Transfer-Learning model outperforms Direct-Specific as well as General-Person (the DPM model lacks the action-specific person detection) on 24 out-of 40. I think it needs to be different from other special postures to be better distinguished, such as brushing teeth and playing the violin. Otherwise, the more complex Transfer-Learning model will learn error information, which not as good as the General-Person model.

The comparison with the SOTA model strongly convinces potential users to advocate the proposed method. After comparing with Action RCNN, Det RCNN, Action-Det RCNN, HOG, CN-HOG with Transfer-Learning on PASCAL VOC 2012, it is clear that Transfer-Learning has the highest mean AP score, which shows that Transfer-Learning can best recognize the actions in static pictures. However, the result after comparing with faster-RCNN and Yolo remains unknown. And which is the best method still needs to be convinced by experiment. Next, by comparing Transfer-Learning and Direct-Specific on Sanford-40, authors put forward another attractive advantage of Transfer-Learning, which can reduce localization errors and confusion with background errors by 4%, respectively. Unfortunately, the author avoids those categories in which Transfer-Learning performs less well. Originally, this should also be a good opportunity to trace back to why they performed in general.

Lastly, we discuss *action classification*. Authors also intuitively compare Transfer-Learning with a series of SOTA to determine the excellence of Transfer-Learning in action classification. On PASCAL VOC 2012, Transfer-Learning has the highest mean AP of 77.0% among MDF, RMP, WAB, Action-RCNN, Wholes and Parts, Action Poselets, Stanford, and Oxford. However, the single indicator makes me worry that it can not well describe classification errors. The lack of a confusion matrix or more specific analysis of the causes of misclassification is also regrettable.

4. Conclusions

In general, the authors apply action-specific person detectors to drive action detection in still images. By posing action detection as a transfer learning, even with the limited training examples, the model works well on Stanford-40 and PASCAL VOC 2012. The capture of human posing and the transfer learning are really interesting ideas, which reflect the improvement of the model on mean AP. However, the reasons for the decline of the classification accuracy of some categories need to be found so that the model can be optimized.

What's more, more experiments could be carried out. For methods, both acceleration and deep learning can optimize the DPM model, data augmentation can be compared with transfer learning, and the updated deep learning model can replace the CNN model in this paper, etc. For evaluation, new data sets with more complex occlusion and background can be tested, and more evaluation parameters can be used, especially the confusion matrix.

References

[1] Static gesture recognition,[Online]. Available:https://help.aliyun.com/document_detail/206650.html/.

[2] Limb movement recognition,[Online]. Available:https://www.zeewain.com/techs/body_movement_detect.

[3] V. Choutas, P. Weinzaepfel, J. Revaud and C. Schmid, "PoTion: Pose MoTion Representation for Action Recognition," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7024-7033, doi: 10.1109/CVPR.2018.00734.

[4] R. Girshick, "From rigid templates to grammars: Object detection with structured models," Ph.D. dissertation, Dept. Comput. Sci., Univ. Chicago, Chicago, IL, USA, 2012.

[5] J. Yan, Z. Lei, L. Wen and S. Z. Li, "The Fastest Deformable Part Model for Object Detection," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2497-2504, doi: 10.1109/CVPR.2014.320.

[6] Y. Tang, X. Wang, E. Dellandréa and L. Chen, "Weakly Supervised Learning of Deformable Part-Based Models for Object Detection via Region Proposals," in IEEE Transactions on Multimedia, vol. 19, no. 2, pp. 393-407, Feb. 2017, doi: 10.1109/TMM.2016.2614862.

[7] Rodrigo Benenson, Mohamed Omran, Jan Hendrik Hosang, Bernt Schiele. "Ten Years of Pedestrian Detection, What Have We Learned?" 2014, european conference on computer vision.

[8] R. Gao, B. Xiong and K. Grauman, "Im2Flow: Motion Hallucination from Static Images for Action Recognition," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018 pp. 5937-5947. doi: 10.1109/CVPR.2018.00622.

[9] T.He, Z. Zhang, H. Zhang and Z. J. Mu, "Bag of Tricks for Image Classification with Convolutional Neural Networks",2018