WEAKLY SUPERVISED HUMAN BODY DETECTION UNDER ARBITRARY POSES

Yawei Cai, Xiaoyang Tan

Nanjing University of Aeronautics and Astronautics
Department of Computer Science and Technology
x.tan@nuaa.edu.cn

ABSTRACT

In this work we study the problem of weakly supervised human body detection under difficult poses (e.g., multi-view and/or arbitrary poses) within the framework of multi-instance learning (MIL). We first point out the existence of the so-called "vanishing gradient" problem in MIL with a noisy-or rule as its bagging model. This is mainly due to the independence assumption of the noisy-or rule, which significantly reducing the magnitude of gradient under a weak initial instance-level model. To address this issue, we propose an iterative selective MIL method in which 1) the noisy-or rule is replaced with the max rule and only a few instances are included for MIL learning for each bag and for each time, and 2) prior knowledge about the positive instances in terms of few fully supervised samples are employed to improve the robustness. The method is shown to outperform the previous state-of-the-art methods by over 20.0% in accuracy. Finally, we present a new large-scale data set called MPHB (Multiple Poses Human Body) for human body detection under arbitrary poses.

Index Terms— human body detection, weakly supervised learning, multiple instance learning

1. INTRODUCTION

Multi-pose human body detection has many important applications in practice [1]. For example, in human gesture estimation [2][3][4], one often need to detect the position of the human body first to provide a reference location for other human parts, such as head, hands, feet, and so on. For an intelligent robot, its actions should avoid colliding with people, but people in a room will not always be upright - they can be bending, sitting, lying, or in other poses (see Fig.1 for some illustration of human body in different poses). Hence detecting human body under arbitrary poses becomes necessary [5].

However, this topic has not received enough attention yet. One of the most studied topic related to this is the pedestrian detection [6], which has wide applications especially in the urban intelligent transportation system. Nowadays its performance has reached a high level [7], providing valuable lessons in many aspects (e.g., feature extraction, model selection, evaluation methods) to multi-pose human body detection. But pedestrian detection by itself concerns mainly about human body in upright positions.

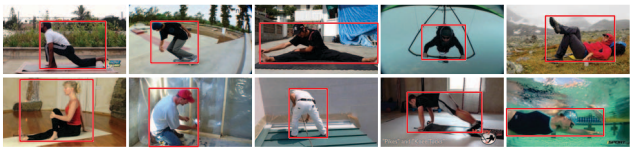

Fig. 1. Illustration of human body under different poses.

There are several challenges for human body detection under arbitrary poses. Besides large appearance variations, various poses make the job of manual annotation difficult and laborious, and usually only weak annotations are available. Multi-instance learning (MIL) is a good tool to address these challenges as it relaxes the requirements for accurate labeling. With MIL, we are even not necessary to annotate the ground truth but just to label whether there is an object of interest in the image.

However, due to the lack of strong supervision information, previous MIL methods have to collect a large number of instances to guarantee that at least one positive instance appears in a positive bag. Unfortunately, doing this not only significantly increases the computational costs, but also is essentially not consistent with the noisy-or bagging model of MIL - we first observe that too many negative instances in a positive bags will hurt the performance in this paper, in the sense that they significantly reduce the magnitude of gradient especially under a weak initial instance-level model. To address this issue, we propose an iterative selective MIL method by replacing the commonly used noisy-or rule with the max rule and by incorporating the prior knowledge about the positive instances regarding few fully supervised samples.

In the field of general object detection, there exist many large-scale public data sets for researchers to use such as Pascal VOC 2007 [8], ILSVRC 2010 [9], ILSVCR 2012 [10]. But these are not quite suitable for human body detection since in these data sets, the person is only one of many categories and its number of images is usually not big enough.

There also have been datasets for human pose estimation like FLIC [11], LSP [12] and MPII Human Pose [13], but they lack annotations of the position or bounding box of human beings needed for performance evaluation. To fill this gap, we annotated a new data set specific to human body detection, named MPHB (Multiple Poses Human Body), containing 26,675 images and 29,732 human bodies, with every human body in each image annotated with a bounding box (Fig. 1). This dataset will be made freely available for research purpose.

The remaining of this paper is organized as follows, after briefly introducing the related work of general object detection in Section 2, we give the description of our algorithm in Section 3 and the data set MPHB in Section 4. Experiments on multi-pose human body detection are presented in Section 5 and we conclude the paper in Section 6.

2. RELATED WORK

There are relatively few works devoted to human detection under arbitrary poses. But one of the most influential method in general object detection is the R-CNN method proposed by Girshick et al. [14], in which a selective search algorithm is introduced to generate sets of candidate detection proposals and these are subsequently fed into a deep network for further verification. The method effectively bypasses the time-consuming sliding window scanning and take the advantage of deep learning, achieving the state of art results on many object detection tasks [15]. A large number of supervised training samples are needed to prevent its complicated likelihood model from overfitting.

In many practical applications, it is difficult or laborious to obtain so many labeled samples or to annotate them manually, and hence people want to localize objects with minimal supervision. This inspires the interest of weakly supervised learning in recent years. Multiple-instance learning (MIL) is a classic approach for this. Recently Cinbis et al. [16] proposed a MIL approach for object detection in which a multi-fold MIL procedure is proposed to prevent training from prematurely locking onto erroneous object locations. Our method is closely related to this but their work does not aim for human body detection under arbitrary poses and their training object is different from ours as well.

3. SELECTIVE WEAKLY SUPERVISED DETECTION

3.1. A Brief Review on Multi-instance learning

In the setting of multi-instance learning [17], we are given N bags of data, denoted by $D = \{x_i, t_i\}_{i=1}^{N}$, where x_i is the $i-$th bag and t_i is its corresponding label. A bag usually contains a set of M instances, i.e., $x_i = (x_{ij})_{j=1}^{M}$, and they are treated as a whole unit in MIL training. Only if there exists no positive instance in a bag can this bag be called negative ($t_i = 0$), otherwise it is a positive bag ($t_i = 1$). Under this definition, we are uncertain which instance is really positive in a positive bag, which may impose a significant challenge on a MIL algorithm. The general MIL algorithms, i.e., multi-instance logistic regression [18] and multi-instance AdaBoost [19] are not robust enough against the noisy data in a positive bag.

In the context of an object detection problem, each image is usually treated as a bag, and each candidate window as an instance in this bag. When testing, each candidate window (instance) has to be assigned a label, which seems to be directly contrary to the definition of multi-instance learning in that no label will be given to a specific instance. One can get around this issue either by using the instance-level model (if exists) for prediction or by thinking of each test instance as a bag consisting of only one single instance and making the prediction using the bag-level model.

Formally, denote the probability of an instance x_{ij} being positive as p_{ij}. To estimate the conditional probability p_i at the bag level, one can fuse the probability at the instance level using different strategies, and two of the most commonly used ones are the max pooling and the noisy-or model, respectively

Max Pooling:

$$p_i = \max_j \{p_{ij}\} \tag{1}$$

Noisy-or:

$$p_i = 1 - \prod_j (1 - p_{ij}) \tag{2}$$

The max pooling strategy aims to find the instance most likely to be positive in a bag but does not care about the labels of other instances. On the contrary, the noisy-or model takes all instances into account but assumes that they are independent of each other.

3.2. The Vanishing Gradient Problem of MIL

Due to the lack of strong supervision information, a common practice in previous MIL detection methods is to collect a large number of instances in a single image to justify the MIL assumption that at least one positive instance appears in a positive bag. But too many negative instances in a positive bag will actually hurt the performance if a noisy-or rule is used. To see this, suppose that we have 1,000 negative instances in a positive bag and each of them are properly assigned a low positive probability of 0.1 by a normal not-so-accurate instance-level model, then according to the noisy-or rule, roughly we would have a probability of $(1 - 0.1)^{1000} \sim 1.7 * 10^{-46}$ that the bag is being negative, no matter how the model assigns posterior probability to the remaining positive instance. This is fine if the p_i is simply used for bag comparison, but it will raise a big problem if we use this for model training, as explained below.

Let's take the MIL logistic regression method (MIL-LR [18]) to illustrate this. It is a linear classifier with its instance level classifier as $y_{ij} = w^T x_{ij} + b$, where w and b are parameters to be learned, and the probability p_{ij} of an instance being positive is modeled using a sigmoid function. To train the model, a negative likelihood is used as loss function,

$$J = -\sum (t_i \ln p_i + (1 - t_i) \ln(1 - p_i)) \tag{3}$$

with its gradient as,

$$\frac{\partial J}{\partial w} = \sum \frac{\partial J}{\partial p_i} \frac{\partial p_i}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial w} \tag{4}$$

Here, $\frac{\partial J}{\partial p_i} = -\frac{t_i}{p_i} + \frac{1-t_i}{1-p_i}$ and $\frac{\partial p_{ij}}{\partial w} = (1 - p_{ij}) p_{ij} x_{ij}$.

Note that when parameter learning, the gradient that evaluates how the small change of instance-level probability p_{ij} influences the bag-level probability p_i (i.e., $\frac{\partial p_i}{\partial p_{ij}}$) plays an important role. Particularly, under the noisy-or rule, this equals to $(1-p_i)/(1-p_{ij})$. Now, let us optimistically assume that we have a small number of negative instance in a positive bag (e.g., only five), each of which will be initially assigned a p_{ij} about 0.5 (since initially we usually only have very weak instance level classifier available), and $p_i = 0.9688$ according to the noisy-or rule. In other words, even with this very weak initial model, the noisy-or rule would give us a very high probability that the bag being positive. Consequently, this will mislead the training by significantly reducing the magnitude of the gradient. We call this the MIL version of the notorious vanishing gradient problem.

To address this issue, in this work we propose to use the max pooling model instead. It depends only on the winner instance x_{ij^*}, and $\frac{\partial p_i}{\partial p_{ij^*}} = 1$, where $p_{ij^*} = \max_j \{p_{ij}\}$. This property makes the model very robust against the inaccurate annotations, if the linear model is reasonable.

3.3. Selective Weakly Supervised Detection

To improve the accuracy of the initial instance-level classifier, we propose a simple Selective Weakly Supervised Detection (SWSD) method that takes advantage of a small amount of supervised information usually available in practice. These few initial bags with fully supervised instances are kept all the time through the whole training procedure to play the role of regularization for a MIL model.

The whole training process is given in Algorithm 1. There are two parameters of this algorithm, i.e., the number of iterations T and the proportions r of samples to be selected at each iteration. Both parameters are set according to cross validation. We may also set a shrinking rate η empirically at each iteration to reduce the size of the searching region, as our detection confidence is expected to increase when more information is incorporated with the training ongoing. In addition, we use semi-random proposals selection as a mechanism to inject the diversity into our detector training, which effectively keeps the algorithm from being trapped in bad local minimum (wrong locations but with high score).

Algorithm 1 Selective Weakly Supervised Detection

1. Initialize training set S_0: fully supervised samples consisted of M ground-truth and M negative instances in positive bags and N empty positive training bags; detection proposals from each image.
2. Train a MIL instance-level classifier using S_0 and locate a relatively large searching region l based on the detection scores of each proposal.
3. For iteration t = 1 to T
   (a) Randomly select a r proportion of instances/proposals in the searching region l from each positive bags.
   (b) Run the current MIL detector over those random instances/proposals and assign a score for each of them.
   (c) Select n top ranking proposals and combine them with the S_0 set to construct a new training sets S_t; locate the next searching region l using these.
   (d) Use S_t to train a new MIL detector.
4. Return the final detector.

4. MULTI-POSE HUMAN BODY DATA SET

Images in our Multiple Pose Human Body (MPHB) data set are selected from LSP [12] and MPII Human Pose [13]. The reasons we select them are two folds: 1)they are originally created for human pose estimation and hence contain mainly persons; 2) they are challenging enough for a person detection task in that they not only include persons in various poses, some other interesting distracting factors such as partial occlusions, dim light, cluttered background, persons in various scales, and multiple persons in one image are not uncommon as well. However, they are ready to be directly used for human body detection due to the lack of necessary ground truth (e.g., human bounding boxes).

We selected 2,000 images from LSP and 24,675 from MPII Human Body, leading to 26,675 images in total. We annotate each image for each subject with bounding boxes, resulting in 29,732 human bodies in all. Among them, about 70% annotations' size ratio (relative to image size) is less than 10%. Fig. 2 gives some illustration of the samples.

We further divide these images into six categorizations according to the poses of persons, i.e., bent (e.g., stooping down or lunges), kneeling, lying (e.g., sleeping or swimming), partial occlusion, sitting and standing upright. These are illustrated in Fig. 2.

We design an evaluation protocol by dividing these images into three partitions. Particularly, the number of images (human bodies) are respectively 8,385(9,732), 8,110(8233), 10,180(11,767) for training set, validation set and test set. Following common practice in this field, when testing we calculate the IoU (Intersection-over-Union) for each test image

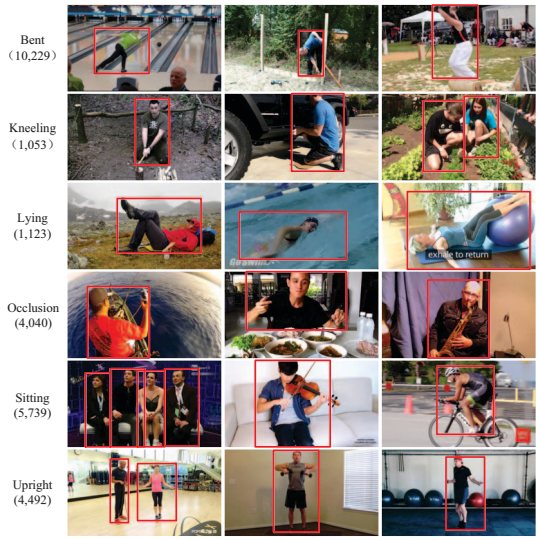

Fig. 2. Illustration of the images in the MPHB datsets, with different types of poses.

and instances whose IoU (Intersection-over-Union) with the ground-truth is more than 0.5 are considered as positive. The final performance is summarized using the average precision (AP) metric.

5. EXPERIMENTAL RESULTS

We use Selective Search method [20] to generate detection proposals and the VGG net [21] for feature representation. In training initially we use 100 fully supervised instances as priors, and the algorithm parameter of r is set to be 5% and T set to be 6 consistently throughout the experiments.

We compare our method with two states of the art weakly supervised object detection methods, i.e., Cinbis et al.'s Multi-fold MIL [16] and Bilen et al.'s Posterior Regularized Latent SVM method [22]. Both algorithms are implemented by us and we have separately tested our implementations by running them on the Pascal VOC 2007 dataset [8], finding that both achieve similar results to those reported in their original papers. To illustrate the contribution of the supervised regularizer, we also evaluate two variants of the proposed SWSD method: the first one is simply to use the initial detector trained with 100 fully supervised instances for testing, and the second is the SWSD algorithm without using any supervised information.

Table 1 gives the results. It can be seen that the proposed SWSD method performs the best among the compared ones. As shown by the table, the proposed method significantly outperforms two previous methods by 18.8% and 24.6% respectively. This indicates that our method is much more robust than the compared methods when detecting highly deformable human bodies.

Method	AP(%)
Posterior Regularized Latent SVM method [22]	10.97
Multi-fold MIL [16]	16.61
SWSD(Ours)	**35.37**
a) Fully supervised detector (with 100 samples)	27.44
b) SWSD without supervised regularizer	21.26

Table 1. Performance comparison (AP %) of various weakly supervised person detection method on MPHB.

Pose	Bilen et al .'[22]	Cinbis et al .'[16]	SWSD
Bent	11.61	14.41	20.68
Kneeling	15.33	15.91	22.52
Lying	7.30	9.53	10.68
Occlusion	13.94	15.61	47.20
Sitting	11.93	17.18	29.63
Upright	25.53	27.50	48.28

Table 2. Performance comparison (AP %) on different poses.

Table 1 also reveals that our performance advantage is preserved even when no supervised regularizer is adopted - in this case, our algorithm achieves an AP accuracy of 21.26%, much higher than the compared ones (10.97% for [22] and 16.61% for [16]). Although this is lower than the fully supervised method with only 100 samples by about 6.2%, the table shows that our algorithm improved upon the fully supervised method by about 8.0%, indicating that the proposed semi-random iterative MIL learning procedure is effective.

Table 2 gives the detailed performance on various poses. One can see that the non-upright poses such as lying, bent, kneeling are more challenging than others due to the large deformations, although our algorithm achieves the best performance consistently over all the poses.

6. CONCLUSION

In this paper we proposed a novel Selective Weakly Supervised Detection (SWSD) method for human body detection under arbitrary poses. This is a modified MIL method in which the classical noisy-or bagging model is replaced with the max pooling principle and a few fully supervised samples are incorporated as prior knowledge. We also contribute a large-scale data set MPHB for this less-studied topic, on which we show that the proposed method outperforms several previous state-of-the-art weakly supervised object detectors.

7. ACKNOWLEDGEMENTS

This work is partially supported by National Science Foundation of China (61373060).

8. REFERENCES

[1] Dalal N and Triggs B, "Histograms of oriented gradients for human detection," *CVPR*, 2005.

[2] Toshev A and Szegedy C, "Deeppose: Human pose estimation via deep neural networks," *CVPR*, 2014.

[3] Yang Y and Ramanan D, "Articulated human detection with flexible mixtures of parts," *Pattern Analysis and Machine Intelligence*, 2013.

[4] Pishchulin L, Andriluka M, and Gehler P et al., "Strong appearance and expressive spatial models for human pose estimation," *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[5] Buys K, Cagniart C, and Baksheev A et al., "An adaptable system for rgb-d based human body detection and pose estimation," *Journal of Visual Communication and Image Representation*, 2014.

[6] Oren M, Papageorgiou C, and Sinha P et al., "Pedestrian detection using wavelet templates," *CVPR*, 1997.

[7] Yadav R, Senthamilarasu V, and Kutty K et al., "A review on day-time pedestrian detection," *SAE Technical Paper*, 2015.

[8] Everingham M, Van Gool L, and Williams C K I et al., "The pascal visual object classes challenge 2007 (voc 2007) results (2007)," 2008.

[9] Berg A, Deng J, and Fei-Fei L, "Large scale visual recognition challenge (ilsvrc), 2010," 2010.

[10] Deng J, Berg A C, and Satheesh S et al., "Imagenet large scale visual recognition challenge (ilsvrc) 2012," 2012.

[11] Sapp B and Taskar B, "Modec: Multimodal decomposable models for human pose estimation," *CVPR*, 2013.

[12] Johnson S and Everingham M, "Clustered pose and nonlinear appearance models for human pose estimation," *BMVC*, 2010.

[13] Andriluka M, Pishchulin L, and Gehler P et al., "2d human pose estimation: New benchmark and state of the art analysis," *CVPR*, 2014.

[14] Girshick R, Donahue J, and Darrell T et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," *CVPR*, 2014.

[15] Zhang Y, Sohn K, and Villegas R et al., "Improving object detection with deep convolutional networks via bayesian optimization and structured prediction," *arXiv preprint arXiv:1504.03293*, 2015.

[16] Cinbis R G, Verbeek J, and Schmid C., "Multi-fold mil training for weakly supervised object localization," *CVPR*, 2014.

[17] Ray S and Craven M, "Supervised versus multiple instance learning: An empirical comparison," *Proceedings of the 22nd international conference on Machine learning*, 2005.

[18] Xu X and Frank E, "Logistic regression and boosting for labeled bags of instances," *Advances in knowledge discovery and data mining*, 2004.

[19] Zhang C, Platt J C, and Viola P A, "Multiple instance boosting for object detection," *Advances in neural information processing systems*, 2005.

[20] Van de Sande K E A, Uijlings J R R, and Gevers T et al., "Segmentation as selective search for object recognition," *ICCV*, 2011.

[21] Simonyan K and Zisserman A, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] Bilen H, Pedersoli M, and Tuytelaars T, "Weakly supervised object detection with posterior regularization," *BMVC*, 2014.