doi:10.21311/001.39.1.35

Coarse-to-Fine: Cascade Regressors in Pedestrian Detection

Mingyang Liu, Fengting Li*, Pei Sun
College of Instrumentation & Electrical Engineering, Jilin University,Changchun 130061, Jilin, China

Abstract

Hand-design feature and sliding window method have been widely used in object detection and achieved great success in pedestrian detection. However, sliding window method costs too much time on calculation and hand-design features lack of prior information since its unsupervised manner. In this paper, we propose a method for pedestrian detection based on Convolutional Neural Networks (CNN). The pedestrian detection is formulated as a CNN-based regression problem towards body regions. To achieve high precise location, cascade of such CNN based repressors is applied in our approach. The approach benefits at the powerful formulation of CNN based feature and has the advantage of high efficiency which capitalized on regression process instead of mass of candidate classifications. Experimental results show that our approach outperforms the state of the art approaches.

Key words: Image Detection, Computer Vision, Image Match

1. INTRODUCTION

Pedestrian detection is one of the hottest topics in computer vision field and can be useful in many commercial applications. But there also are many challenges, such as multiple views, variations of human body poses, occlusions and lighting changes, by which difficulties in pedestrian detection suffered. In past decades, human-designed visual features and sliding window methods are most widely used to handle such problems, however, due to lack of power for information formulation, it's unlikely to solve them well for visual features such as SIFT, HOG, and so on. Besides, since it cost too much on calculation, sliding window methods can hardly be practically applied (M. Mathias, R. Benenson, R. Timofte, 2012).

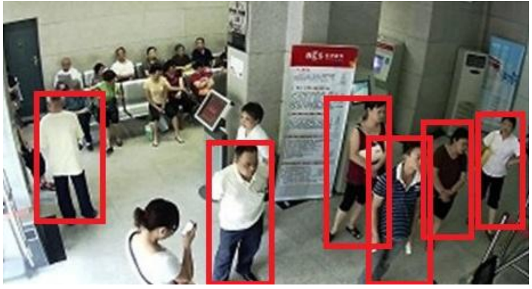

Figure 1. To detect pedestrian in the wild image could be a challenge task, since the multiple views, variations of human body poses, occlusions and lighting changes

To describe human body better, in this work CNN based feature is selected. We utilize recent developments of deep learning techniques and propose a novel algorithm based on a Convolution Neural Network (CNN). CNN has been proved that it outperforms most of other algorithms in high level classification tasks and also can achieve high performance in object detection area. However, it is still a not-answered question whether applying CNN for precise localization of pedestrian. In this paper, we'll try to give a complete solution for these problems and present a simple but powerful formulation of pedestrian detection as a CNN based feature.

In this paper, the pedestrian detection is formulated as a joint regression problem. CNN based feature is extracted on the input full image which should be resized to a fixed size. And the location of pedestrian is regressed to using the classical 8-layerd CNN features. There are several advantages of this formulation. First, the CNN feature is powerful enough to capture all kinds of variations of human bodies. Second, it will be much more efficient to get human body location by a capable regressor such as SVR, etc than to get a scalable bounding box after a heavy confidence calculation for a mass of candidate sub-windows. Regression-like method makes the precise pedestrian detection an achievable and practical approach for many real-time applications (L. Bourdev and J. Brandt, 2005).

Furthermore, we propose a cascade of CNN-based pedestrian detectors. Such a cascade regressor makes much more precise location prediction for human bodies. Taken the coarse body locations as input, CNN-based

SVR regressors which refine the detection results could be learned.Via the benchmarks of pedestrian detection, we show that out approach outperforms the state-of-art of all reported results.

The rest of this paper will be organized as follows: firstly, related work on such field will be introduced in section 2; and we will describe details of our approach in section 3; experimental result can be found in section 4, and finally we will give our conclusion in section 5.

2. RELATED WORK

Hand-designed feature is a traditional topic which improves detection performance directly. In the past decades, the most widely used visual features include Haar-like features, scale-invariant feature transform (SIFT) (L. Bourdev, S. Maji, T. Brox, and J. Malik, 2010), dense SIFT, color SIFT, histogram of gradients (HOG) (Alexander Toshev Christian Szegedy, 2014), gradient histogram, local binary pattern (LBP) and so on. And other low-level features such as color histogram, normalized gradient, and MSER also improve performance of pedestrian detection.

In most cases, pedestrian detection is considered as a classification task by scanning an image with sliding windows in multiple scales. There are a lot of approaches developed based on the method mentioned above, such as calculating the confidence of a window enclosing a human body with a probability prediction model, and the discriminative classifiers, such as boosting classifiers and SVM, trying to learn the parameters to distinguish positive and negative sub-windows.

Part-based models have achieved great success in object detection and recognition (N. Dalal and B. Triggs, 2005). Furthermore, the deformable part based model (DPM) is capable to detect objects with some variations such as pose change, appearance distortions and so on. And to make the process speed up, cascade-based DPM is developed and show far better efficiency. Multiple components are able to handle the pose changes.

Recently, deep learning based methods have been successfully applied in most computer vision fields such as hand written digit recognition, object segmentation, face recognition, scene classification, object detection and recognition, and so on. In most cases, deeper architectures gain better performance than shadow ones.

The closest work to ours uses convolution NNs as visual feature with selective search algorithm to generate proposals, and followed by a heavy svm classifier to separate the positive and negative ones among the proposals, SVR model based on the same feature is used to refine the initial location of a certain object (N. Dalal, B. Triggs, and C. Schmid, 2006). However, this work does not apply a cascade of refine models, besides, classification from the proposals demands CNN-based feature extraction for each one of the proposals, which brought a heavy burden for an efficient calculation.

3. CALCULATION

We have to give some notations first. To describe a pedestrian location, we encode the coordinates defined as $\mathrm{Y} = (\mathrm{Y}_{\mathrm{tl}}, \mathrm{Y}_{\mathrm{br}})$, where Y_{tl} contains x and y coordinates and denotes the top-left corner of a bounding box, and Y_{tl} contains x and y coordinates of the bottom-right corner. And an input image can be described as (X; Y), where x stands for the image data and ground truth location can be expressed as Y (Desai and Ramanan, 2012).

Besides, as the bounding box coordinates are in absolute, it may suffer huge variation when scalar of input image changes. As a result, normalizing the location should be a proper method to improve accuracy. In most cases, the coordinates can be normalized by dividing by the width or height of the given image. As the bounding box is formulated as a center pointb_{c}, width b_{w} and heightb_{h}: $\mathrm{b} = (\mathrm{b}_{\mathrm{c}}, \mathrm{b}_{\mathrm{w}}, \mathrm{b}_{\mathrm{h}})$, we could formulate the location Y as the translation of center and the scalar of the box size as follows:

$$\mathrm{N}(\mathrm{Y},\mathrm{b}) = \begin{pmatrix} \frac{1}{b_w} & 0 \\ 0 & \frac{1}{b_h} \end{pmatrix} (Y - b_c) \quad (1)$$

3.1. CNN-based Proposal selection

To get the best initial location of the entire detection procedure, we've to select the proposals first. In this paper, instead of using multiple-scale sliding window to generate mass of sub-windows, we use prior information to generate proposals, as in our approach; there is no need to initiate too much precisely. In this paper, K-Means-like method is applied to generate proposals (J. Xiao, 2012), we use a 4-dimensioned vector to describe each training sample as follows:

$\mathrm{P} = (\mathrm{c}_{\mathrm{x}}, \mathrm{c}_{\mathrm{y}}, \mathrm{r}_{\mathrm{w}}, \mathrm{r}_{\mathrm{h}})$,where c_{x} denotes x coordinate of the center of a labeled bounding box, and c_{y} y coordinate, normalized by dividing the width and height of the given image, and r_{w} andr_{h} indicates the normalized rectangle width and height of a labeld bounding box. In our approach, K of the K-Means method is set to 10, to achieve a trade-off between effectiveness and efficiency.

SVM is applied to score each proposal, from which CNN-based feature is extracted. This feature will be described in details in the coming section, as it's just the same as the feature used in regression procedure. Only the proposals gaining a score higher than a fixed threshold could be considered to be the candidates of coming regression module (Doll´ar, Appel, and Kienzle, 2012).

3.2. Pedestrian detection as CNN-based Regression

In this paper, the problem of pedestrian detection is treated as regression; where the we train a regression model which for the input image X regress to a normalized location vector. Thus, according to the normalization transformation of Eq. (1) the bounding box location in absolute image can be described as follows:

$$Y^{*} = N^{-1}(\varphi(N(x);\theta)) \ (2)$$

As can be seen in the figure, instead of predicting the absolute location of the input image, displacement from the initial proposals is predicted, for the reason of which could look more into the details of the regression process. The displacement could be formulated as follows:

$$\Delta x = (b_x - g_x)/b_w (3)$$

$$\Delta y = (b_y - g_y)/b_h (4)$$

$$\Delta w = \log(b_w/g_w) \ (5)$$

$$\Delta h = \log(b_h/g_h) \ (6)$$

Although the formulation is simple, the power to solve complex problem of detect pedestrian in wild scene is derived from the Convolutional Neuron Network (CNN) based feature. Such a convolutional network contains a number of layers, each of which consists of a linear transformation and a non-linear one. It takes an image of predefined size as input for the first layer, and the input vector has a size three times to the number of pixels (Doll´ar, Belongie, and Perona, 2010). The 6th fully connected layer (FC6) is selected as the CNN-based feature, it has been proved that the FC6 contains the most information which benefits the object detection task and other related applications. And the target values of the regression will beoutput by SVR regression models. In our case the center point and scalar factors of a bounding box totally four target values are regressed, so that four SVR models should be trained.

We base the architecture of our CNN feature on the work by Zeiler for image classification for its outstanding results on object detection. In the training stage, the network consists of 8 layers, of which C denotes a convolutional layer, LRN stands for a local response normalization layer, and P indicates a pooling layer and F a fully connected layer. Only parameters belong to C and F layers need to be learned, while there is no parameter in the rest of layers (Doll´ar, Tu, Perona, and Belongie, 2009). In convolutional layers, the size of parameters is defined as (width, height, depth), where the first two dimensions define the spatial shape of a given convolutional kernel while the depth gives the number of filters. Both in C and F layers, the learned parameters formulate a linear transformation which always is followed by a nonlinear one. In our work, rectified linear unit (RELU) is selected for the nonlinear transformation.

The architecture of the entire network can be described as Figure 2:

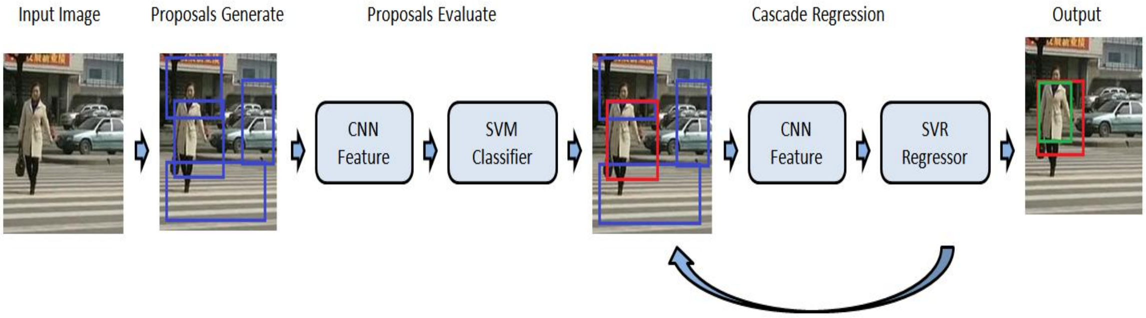

Figure 2. Framework of our proposed cascade regression approach

In this paper, to accelerate the detection procedure, we modify structure of Zeiler Net slightly. First, the size for the first C layers is 7x7 and the number of filters remains 96, the size of the second one still remains 5x5 and the other three are3x3 as origin, but for the middle three layers the numbers of kernels is set to be 192. Both the reduction of size and number is for speeding up of our feature extraction procedure as we try to achieve a real-time detection system. Pooling is applied after three convolutional layers and could contribute to increase efficiency due to the reduction of resolution. The input size of an image is also resized to 114x114fedding into the network. The total number of parameters in the above model is about10M (Zhang, Donahue, Girshick, et al, 2014).

The use of CNN-based feature is inspired by its outstanding results on both classification and detection tasks. We will show that such CNN-based feature can be used to learn a model which outperforms the state-of-art in pedestrian detection task (Dollar, Wojek, Schiele, and Perona, 2011).

In this paper, we select the FC6 layer as CNN feature and use it to train our SVR regressors. Additionally, the use of the CNN feature needs to train a specific model for pedestrian detection task. Such a model and the feature are learned from the data which take the pedestrian as ground truth labels. All the internal features are shared by all regressors.

3.3.Cascade of CNN-based Regression

The regression method from the previous stage has the advantage that the translation estimated based on the full image and thus relies on context. However, due to the variation of the distribution in a given image of the pedestrian, the network has limited capacity to be localized to the exact precise location at one time. For the first time of regression, it just learns regressors capturing location information at coarse scale. In order to achieve better precision, cascade of regressors is proposed. At the first stage, the cascade begins by estimating a coarse displacement from the proposals outputted by the first proposal selection procedure (Andriluka, Roth, and Schiele, 2009; Dantone, Gall, Leistner, and Van Gool, 2013). At subsequent stages, additional CNN regressors are trained to predict a displacement of the previous stage to the true location. Meaningfully, each subsequent stage could be considered as a refinement to the current location.

4. EXPERIMENTAL RESULT

4.1.Setup

There are large number of benchmarks for pedestrian detection. In this paper the datasets which have large number of examples sufficient for training a large model such as the proposed DNN, and are realistic and challenging are used.

We run the proposed approach on three public datasets, the Caltech dataset, the ETHZ dataset and the TUD-Brussels dataset. Only the reasonable subset, i.e. images with the size larger than 100 pixels, non-occluded or partially occluded pedestrians are considered as the testing datasets.

4.2.Metrics

We use log-average miss rate as the metrics for our experiment. The log-average miss rate could be formulated as the average of nine False-Positive-Per-Image (FPPI) rates evenly spaced in log-space in the range from0.01 to 1. We use the metric to evaluate the overall performance as suggested in (P. Dollar, C. Wojek, B. Schiele, and P. Perona, 2011).

4.3.Experimental Details

As we used the architecture of cascade models, training procedure also need to be divided into multi-stages. Firstly, we train the mentioned Zeiler-Net on Imagenet Public classification dataset which includes 1000 categories and totally about 1 million images. Although the training set is not closely related to our pedestrian detection task, since its wide distribution on wild scenes, features trained on this dataset naturally have the ability for of generation on various tasks.

For the cascade regressor training, only the proposals gaining a SVM score higher than 0.5 are considered to be candidates, and the output of previous stage is used as the input of the next stage training. In our approach, the size of cascade is set to be 3.

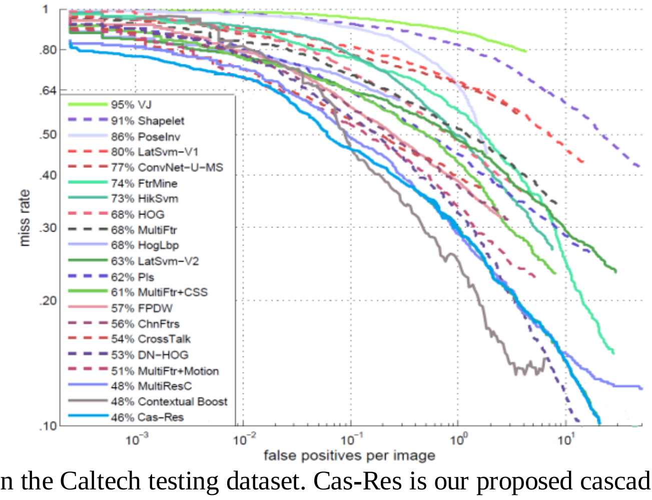

Figure 3. Comparison on the Caltech testing dataset. Cas-Res is our proposed cascade regression method.

4.4.Resultss

We use the training set in the provided datasets as training data, and test on the testing set. Figures 3 to 5 show the experimental results. The compared approaches list their detection results in the opening benchmark website “www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/”. Experimental results show that our approach outperforms the state-of-art in this benchmark.

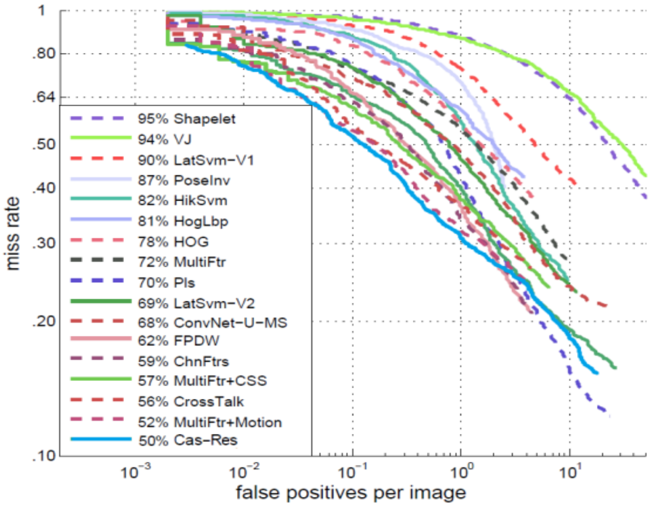

Figure 4. Comparison on the TUD testing dataset. Cas-Res is our proposed cascade regression method.

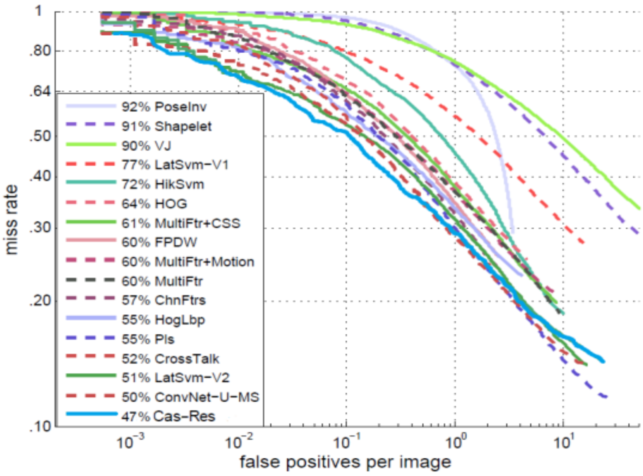

Figure 5. Comparison on the ETHZ testing dataset. Cas-Res is our proposed cascade regression method.

5. CONCLUSIONS

In this paper, we present an approach of Convolutional Neuron Network (CNN) to pedestrian detection. The proposed method of considering the problem to be a CNN-based regression to displacement of coordinates and the presented cascade of such regressors shows advantages on capturing context details rather than the holistic information. As a result, we achieved state-of-art or better results on several challenging academic datasets.Besides, we show that using a fine-tuned convolutional neural network, which was pre-trained for classification tasks and fine tuned on labeled dataset for pedestrian detection, could improve the detection performance significantly.In future, we plan to focus on novel architectures which could potentially improve localization performance further and in pedestrian detection in particular.

Acknowledgements

This work has been supported by Scientific Development Funding of Jilin Province(No:20130305020GX).

REFERENCES

Alexander Toshev Christian Szegedy (2014) “DeepPose: Human Pose Estimation via Deep Neural Networks”, *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, p.p.1653-1660.

C. Desai and D. Ramanan (2012) “Detecting actions, poses, and objects with relational hraselets”, *Lecture Notes in Computer Science*, 7575, p.p.158-172.

J. Xiao (2012) “Contextual boost for pedestrian detection”, *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, Providence, p.p.2895-2902.

L. Bourdev and J. Brandt (2005) Robust object detection via soft cascade”, *Proc. of Intl. Conf. on Computer Vision and Pattern Recognition*, p.p.236-243.

L. Bourdev, S. Maji, T. Brox, and J. Malik (2010) “Detecting people using mutually consistent poselet activations”,*Proc. of European Conference on Computer Vision*, p.p.168-181.

M. Andriluka, S. Roth, and B. Schiele (2009) “Pictorial structures revisited: People detection and articulated pose estimation”, *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, p.p.1014-1021.

M. Dantone, J. Gall, C. Leistner, and L. Van Gool (2013) “Human pose estimation using body parts dependent joint regressors”, *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, p.p.3041-3048.

M. Mathias, R. Benenson, R. Timofte, and L. (2012) “Van Gool. Pedestrian detection at 100 frames per second”, *Proc. ofIEEE Conference on Computer Vision & Pattern Recognition*, p.p.2903-2910.

N. Dalal and B. Triggs (2005) “Histograms of oriented gradients for human detection”, *Proc. ofIEEE Conference on Computer Vision & Pattern Recognition*, p.p.886-893.

N. Dalal, B. Triggs, and C. Schmid (2006) “Human detection using oriented histograms of flow and appearance”, *Proc. of European Conference on Computer Vision*, p.p.428-441.

P. Doll´ar, R. Appel, and W. Kienzle (2012) “Crosstalk cascades for frame-rate pedestrian detection”, *Proc. ofEuropean Conference on Computer Vision*, p.p.645-659.

P. Doll´ar, S. Belongie, and P. Perona (2010) “The fastest pedestrian detector in the west”, *Proc. ofBritish Machine Vision Conference, Aberystwyth*, p.p.1-11.

P. Doll´ar, Z. Tu, P. Perona, and S. Belongie (2009) “Integral channel features”, *Proc. of British Machine Vision Conference*, p.p.1-11.

P. Dollar, C. Wojek, B. Schiele, and P. Perona (2011) “Pedestrian detection: An evaluation of the state of the art”, *IEEE Trans. on Software Engineering*, 34(4), p.p.743–761.

Zhang N, Donahue J, Girshick R, et al. (2014) “Part-Based R-CNNs for Fine-Grained Category Detection”, *Proc. of European Conference on Computer Vision*, p.p.834-849.