ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 155

2018 International Conference on Advanced Control, Automation and Artificial Intelligence (ACAAI 2018)

Context-aware Pedestrian Detection with Salient Region Self-growing in Far-infrared Images

Hao Sheng[1,2], Meiyuan Liu[1,*], Yanwei Zheng[1,2] and Yang Liu[1,2]

[1]State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, P.R.China

[2]Shenzhen Key Laboratory of Data Vitalization, Research Institute in Shenzhen, Beihang University, Shenzhen, P.R. China

***Abstract*—In this paper, we present a new framework to detect pedestrians in infrared images. The framework consists of a candidate generation module and a classification module, both of which are implemented based on convolution neural network. Specifically, we learned effective segmentation threshold by deep learning methods, and proposed a salient region self-growing algorithm to generate candidates. Besides, we conducted context-aware classification on the candidates to reduce the false positives using cues from the context. We achieved state-of-the-art result on a public dataset, which has shown the effectiveness of the proposed method.**

Keywords—pedestrian detection; infrared; neural network

I. INTRODUCTION

Pedestrian detection is of fundamental importance to computer vision due to its widespread application in modern society such as advanced driver assistant system (ADAS). When helping drivers avoid traffic accidents, ADAS based on visible light camera relies on good illumination condition, as a result of which ADAS based on infrared camera is necessary. The characteristic of infrared imaging—higher intensity of objects in higher temperature—brings up a lot of difficulties for pedestrian detection, such as various appearance of the pedestrians, lack of texture features, low resolution of the images and so on.

Even though, plenty of research work has been done on this field. Current pedestrian detection algorithms on infrared images mainly consist of three procedures: candidate generation, feature extraction and classification. The most common used candidate generation method is intensity segmentation [1], in which threshold values are calculated to separate the pedestrians or just the seemingly brightest heads [3]. Problem is that the threshold values given in these methods are mainly based on observation and experience, thus cannot handle the samples beyond the empirical hypothesis, which are not a few. Besides, the temperature over a pedestrian is not uniform, making it a complicated task to segment the whole pedestrian. And considering the head always being the brightest part of a pedestrian is not a reliable premise.

In the feature extraction step, both traditional and deep learning features have been used, like histogram of oriented gradients (HOG) [4], local binary pattern (LBP) [5], variations of the two, and convolution features [6]. The extracted features will then be classified as pedestrians or non-pedestrians with a machine learning algorithm, such as support vector machine (SVM) [4], AdaBoost [4], sparse representation classifiers [4] and convolution neural network (CNN) [6]. CNN and deep learning has proved its superiority over traditional methods on a lot of computer vision problems, but there still exists a kind of hard negative sample in infrared images, which cannot be correctly detected even with a CNN-based classifier. They come into being because lamps and tires of moving vehicles are in high temperature and constantly changing local environment, which often somehow form a patch highly confusing with the appearance of a pedestrian.

To deal with the problems above, we propose a novel framework for pedestrian detection in far-infrared images. It contains a candidate generation module and a classification module, both of which are implemented based on CNN. For candidate generation, we use a threshold value calculation (TVC) layer, which has powerful study ability, to explore the intensity distribution regularity of pedestrian and learn the threshold. We set a more robust and reliable segmentation objective, and use a salient region self-growing (SRG) layer to locate candidates through the segmented foreground regions. For classification, we implement a context-aware classifier which detects the local context at the same time to promote the performance of the detector. Furthermore, we combine the two modules in a unified network named SCNN.

Contributions of this paper includes:

- A new framework SCNN for end-to-end pedestrian detection in infrared images based on deep learning methods;
- A new candidate generation method for infrared images which is implemented in layers of a neural network;
- A context-aware classifier which can well handle the hard negative samples in infrared image.

We conducted experiments on a public infrared dataset LSI [7] and a self-established dataset HYG. SCNN achieves state-of-the-art result on LSI, and the effect of each part of it is proved on HYG.

II. FRAMEWORK

The overall architecture of SCNN is illustrated in figure 1. In this section, we will introduce the details of: candidate generation module, candidate classification module, and the loss function of the network.

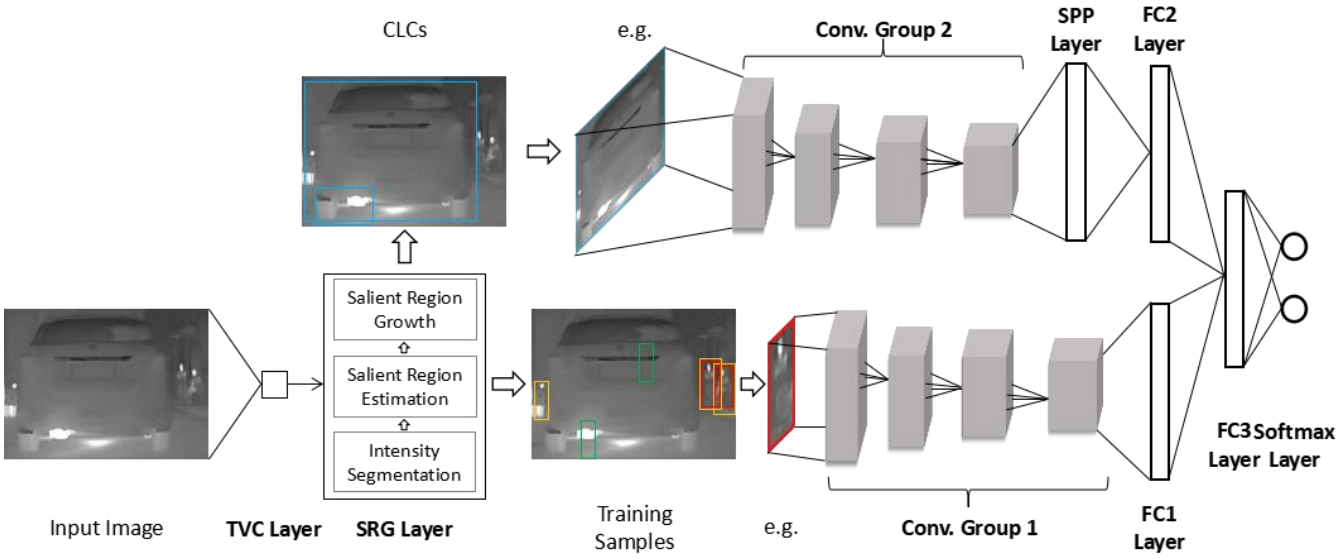

FIGURE I. ARCHITECTURE OF THE PROPOSED SCNN FRAMEWORK.

A. *Candidate Generation Module*

This module includes two layers: the TVC layer and the SRG layer.

1) *TVC layer*

We adopt the convolution operation to be the threshold computation function, which will automatically get optimized in training. Given an input image **I** in size $H \times W \times 3$, TVC layer conducts convolution operation on **I** with a kernel **e** in the same size of **I**. The resultant feature map is a scalar, and can be formulated as:

$$T = I * e, \tag{1}$$

which is right the threshold value we learned.

2) *SRG layer*

Candidates are generated in this layer. Taking the threshold *T* and **I** as inputs, it will first perform binary segmentation on **I.** The segmented foreground regions are defined as salient regions, which should include the brightest body parts of each pedestrian on **I**. Taking a circumscribed rectangle of each salient region as a blob, we are now to obtain corresponding candidates from these blobs.

As aforementioned, the whole body or just head are not optimal segmentation objectives. Instead, we define a constant set $P = \{\text{HD, UB, LB, WB}\}$, representing the head, upper, lower, and whole of the body, as the range of body parts that a salient region can belong to. *P* is a more comprehensive hypothesis of which is the brightest part of a pedestrian body, and the parts in *P* are still valid for we to locate the candidates from them. That is to say, a more elaborate partition, like a hand, will cost much heavier computation to correctly locate the candidates from it, while efficiency is vital to ADAS.

Given a blob *b* with (x_1, x_2) and (x_3, x_4) being its top left and bottom right coordinates ($x_3 > x_1$), we will first categorize it into reasonable body parts in *P* according to its aspect ratio. Once the position of *b* is confirmed, obtaining candidates from it is a matter of deciding the scale for expansion. Due to the symmetry of the parts in *P*, we can develop a uniform set of self-growing rules based on the regularities of structure and proportion of human bodies. Let λ be the horizontal growing coefficient of the left and right sides of *b*, μ and ν are of the top and bottom sides, and $c(y_1, y_2, y_3, y_4)$ is the *k*-th candidate grown from *b*, we can formulate the self-growing rules as:

$$y_1 = [\rho(\mu+1) + (1-\rho)(\nu+1)]x_1 - [\rho\mu + (1-\rho)\nu]x_3 \tag{2}$$

$$y_3 = [\rho(\nu+1) + (1-\rho)(\mu+1)]x_3 - [\rho\nu + (1-\rho)\mu]x_1 \tag{3}$$

$$y_2 = (\lambda+1)x_2 - \lambda x_4 \tag{4}$$

$$y_4 = (\lambda+1)x_4 - \lambda x_2, \tag{5}$$

where

$$\nu = \frac{AR(2\lambda+1)}{ar} - \mu - 1, \tag{6}$$

in which *ar* is the aspect ratio of *b*, and *AR* is the constant aspect ratio of all candidates. Parameter $\rho \in \{0,1\}$ decides which of the top and bottom sides grows first with coefficient μ, and ν is self-adaptive according to μ. Assignment of the coefficients depends on how we define the range of the parts in *P* and how we divide the aspect ratio of *b* to separately discuss. We analyze the issue in experiments and give one of the well-performed example in figure 2:

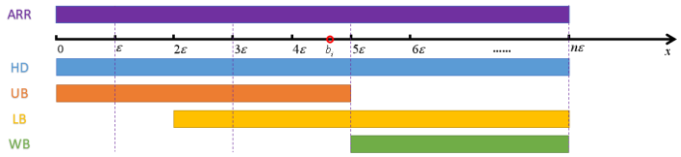

FIGURE II. ASSIGNMENT OF BODY PARTS AND ASPECT RATIO RANGE (ARR)

in which ε=1/3 *and* n=15.

B. *Candidate Classification Module*

This module contains a basic CNN classifier, and a sub-net conducting local context detection to assist the final decision of the candidates.

1) *CNN for classification*

We choose CNN as the base model for the binary classification of the candidates. Convolution group 1 receives candidates in a uniform size of $128\times48\times3$ in our experiment. Number and size of filters at each convolution layer are as: $24\times5\times5\times3$, $32\times5\times5\times24$, $64\times3\times3\times32$, $32\times3\times3\times64$. Feature maps in size $32\times12\times2$ are fully connected to a 256-d feature vector in FC1 layer, which is a higher-order representation of the candidates. We keep the least parameters when the accuracy no longer goes up in training.

2) *Context-aware classifier*

We give an example of a hard negative sample aforementioned in figure 1. In the picture labeled "Training Samples", the leftmost box in yellow frame is a typical one. Enlightened by that human can tell the right answer due to the context, we update the detector by enabling it to perceive the surrounding context. When the foreground is not distinguishable any more, we count on the context-aware ability to improve the performance of our detection method.

We adopt neural network model to extract the contextual features and integrate them with the candidate features. For each candidate sent into the classification module, we attach a piece of local context with it, named candidate with local context (CLC). The classification module splits into two streams here, one takes in the candidates as usual (FEC sub-net), and the other takes in CLCs (FEL sub-net). FEC focus on recognizing the details of candidates such as intensity, texture and shape features, while FEL performs local context analysis and scene detection.

The candidates and CLCs will be represented in feature vectors at FC1 and FC2 layers respectively, both of which are fully connected layers. To utilize the contextual cue in detection, we use FC3 layer to fully connected to the two vectors simultaneously, and learn the feature fusion function through training of SCNN. By constantly optimizing the parameters of this layer over iterations, an ultimate 256-d feature vector comes into being in FC3. It is context sensitive, and again fully connected to two softmax units next layer. Each of the units computes a value indicating the probability of the candidate being a positive sample.

C. *Loss Function*

Although we integrate the candidate generation and classification procedures into SCNN, the loss functions of the two are different from each other.

1) *Candidate generation*

For each input image **I**, there is an ideal threshold T_g by which we can find every pedestrian in **I**. We assign T_g the minimum of the four parts' maximum average intensity values of all the pedestrians in **I**. The L1 distance between T and T_g can measure the loss of T, but not enough. Let $S_c=\{c_1,c_2,...,c_n\}$ be the candidate set of image **I**, $G=\{g_1,g_2,...,g_m\}$ be the ground truth box set. The difference between S_c and G is more intuitive for evaluation, so we use $loss(S_c,G)$ be the penalty term in our loss function.

When g_j has an intersection-over-union (IoU) greater than 0.5 with any of c_i, g_j is considered matched. And the matching extent can be measured by $d(c_i,g_j)$, L2 distance of the feature vectors of the two. $loss(S_c,G)$ adds up when: g_j is missed by all the c_i , and c_i misses all the g_j . The proportion of matched g_j is recall rate, which is the first priority of our optimization objective. We first discuss $loss(S_c,g_j)$:

a) g_j *is matched by at least one candidate:* we let the penalty being little under this situation by taking only the smallest one of all the $d(c_i,g_j)$.

$$loss(S_c,g_j)=\min_{h_j(c_i)>0}\{d(c_1,g_j),d(c_2,g_j),...,d(c_n,g_j)\}, \tag{7}$$

in which $h_j(c_i)=1$ when g_j is matched by c_i and 0 otherwise.

b) g_j *is missed by all candidates:* It will be punished heavily for irreparably lowering the detection accuracy. The penalty burdens every $d(c_i,g_j)$ as:

$$loss(S_{c,}g_j)=\sum_{i=1}^{n}d(c_i,g_j). \tag{8}$$

So we have $loss(S_c,G)=$

$$\sum_{j=1}^{m}(\min\{\sum_{i=1}^{n}h_j(c_i),1\}\min_{h_j(c_i)}\{\{d(c_i,g_j)\}\}+(1-\min\{\sum_{i=1}^{n}h_j(c_i),1\})\sum_{i=1}^{n}d(c_i,g_j)). \tag{9}$$

2) *Candidate classification*

In softmax layer at the end of SCNN, each candidate will be scored of how much the probability that it contains a pedestrian. We use negative log-likelihood to measure the loss

between the estimation and the real labels of the candidates. The average cost for *m* candidates

$$J_{cls} = -\frac{1}{m}\sum_{i=1}^{m}(y_i \log(f(W,b,x)) + (1-y_i)\log(1-f(W,b,x))). \quad (10)$$

III. EXPERIMENTS

A. Dataset and Training

LSI includes 6154 training and 9059 testing images, and HKG includes 17753 and 17005 ones respectively. SCNN is implemented based on Caffe [8]. The training strategy for SCNN is: firstly, using ground truth, shown in red box in figure 1, and the relevant CLCs, in blue box, to train the context-aware classifier; then training the TVC layer using the same data with the parameters of the classifier part being fixed; next using candidates, in yellow box produced in TVC layer, and relevant CLCs, to retrain the classifier; at last we finetune the overall SCNN for further optimization. Each time for training, we randomly select twice as many negative samples as positive ones from the input image, shown in green box.

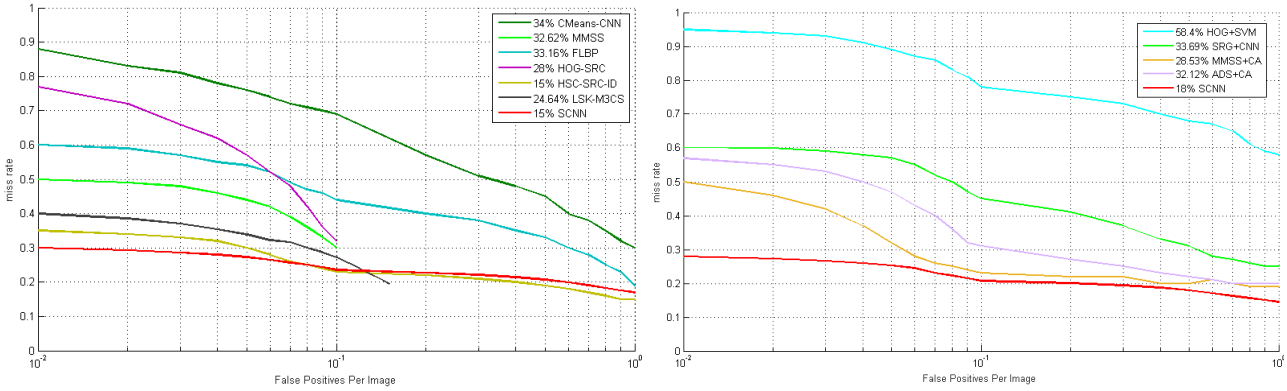

FIGURE III. DET CURVES OF RESULTS ON LSI (LEFT) AND HYG (RIGHT) DATASETS.

B. Evaluation and Results

We follow the evaluation criterion given in [9]. Results are shown in the detection error trade-off (DET) curves of miss rate against false positives per image (FPPI). The percentages in the legend are log-average miss rates (LAMR) presenting the overall accuracy of the methods. The lower LAMR is, the better accuracy the method gets.

The left chart in figure 3 shows the result on LSI of our method comparing with the state-of-the-art method [4] and several other well-behaved methods. Our method SCNN, in red color, achieves 15% LAMR and very close performance with [4], and outperforms it when FPPI is relatively low, owing to the context-aware classifier. The HSC-SRC-ID [4] uses a sparse representation classification. As for other methods, LSK-M3CS [10] uses a linear support tensor machine with LSK channels, achieves 24.64%; HOG-SRC is the old version of [4], achieves 28%; MMSS [2] is a super-pixel segmentation method, achieves 32.62%; FLBP [5] uses the variation of LBP features, achieves 33.16%; CMeans-CNN [6] is also based on CNN, achieves 34%.

The right chart of figure 3 shows the results of several contrast experiments on HYG dataset, which aim at proving the effect of our candidate generation and classification modules. We set a baseline 58.4% with HOG+SVM method, and SCNN reports 18%. Among the contrast experiments, SRG+CNN is SCNN excluding the context-aware part, reports 33.69%; MMSS+CA and ADS+CA replace our candidate generation module with MMSS [2] and ADS [1] while the other operations remain the same as SCNN. The three methods all behave poorer than SCNN to some extent, and the extent shows the effect of the corresponding missing part in SCNN.

IV. CONCLUSION

SCNN reported closely result with the state-of-the-art method on a public dataset, which proves that the framework we propose, a context-aware classifier with SRG algorithm, is effective for pedestrian detection in infrared images.

ACKNOWLEDGMENT

This study is partially supported by the National Key R&D Program of China(No.2017YFC0806500), the National Natural Science Foundation of China(No.61472019), the Macao Science and Technology Development Fund (No.138/2016/A3), the Program of Introducing Talents of Discipline to Universities and the Open Fund of the State Key Laboratory of Software Development Environment under grant SKLSDE-2017ZX-09, the Project of Experimental Verification of the Basic Commonness and Key Technical Standards of the Industrial Internet network architecture. Thank you for the support from HAWKEYE Group.

REFERENCES

[1] G. H. WANG, Q. LIU, “Far-infrared based pedestrian detection for driver-assistance systems based on candidate filters, gradient-based feature and multi-frame approval matching,” Sensors 15(12), 32188–32212 (2015).

[2] N. SUN, F. JIANG, H. YAN, J. LIU, G. HAN, “Proposal generation method for object detection in infrared image,” In Infrared Physics & Technology, Volume 81, 2017, Pages 117-127.

[3] U. Meis, M. Oberlander, W. Ritter, “Reinforcing the reliability of pedestrian detection in far-infrared sensing,” In Proceedings of the IEEE Conference on Intelligent Vehicles Symposium, Parma, Italy, 14–17 June 2004; pp. 779–783.

[4] B. QI, V. John, Z. LIU, S. Mita, “Pedestrian detection from thermal images: A sparse representation based approach,” In Infrared Physics & Technology, Volume 76, 2016, Pages 157-16.

[5] N. SUN, H. YAN, L. LIU, “A filtered local pattern descriptor for face recognition and infrared pedestrian detection,” [J]. (2017-4-19), 2017.

[6] V. John, S. Mita, B. LIU, B. QI, “Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neuralnetworks,” In Proceedings of the IEEE Conference on Machine Vision Applications Proceedings, 2015, pp. 246–249.

[7] D. Olmeda, C. Premebida, U. Nunes, J. Armingol, A. Escalera, LSI Far Infrared Pedestrian Dataset.

[8] Y. JIA, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe:Convolutional architecture for fast feature embedding. arXiv preprint:1408.5093, 2014.

[9] P. Dollar, C. Wojek, B. Schiele, and P.Perona, “Pedestrian detection: An evaluation of the state of the art,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 34:743–761, 2012.

[10] S. K. Biswas, P. Milanfar, “Linear Support Tensor Machine with LSK Channels: Pedestrian Detection in Thermal Infrared Images,” IEEE Transactions on Image Processing, 2017, 26(9):4229-4242.