Pedestrian Detection in Thermal Images Using Adaptive Fuzzy C-Means Clustering and Convolutional Neural Networks

Vijay John
Toyota Technological Institute, Japan
vijayjohn@toyota-ti.ac.jp

Zheng Liu
Toyota Technological Institute, Japan
zhengliu@toyota-ti.ac.jp

Seiichi Mita
Toyota Technological Institute, Japan
smita@toyota-ti.ac.jp

Bin Qi
Toyota Technological Institute, Japan
qibinwinter@gmail.com

Abstract

Pedestrian detection is paramount for advanced driver assistance systems (ADAS) and autonomous driving. As a key technology in computer vision, it also finds many other applications, such as security and surveillance etc. Generally, pedestrian detection is conducted for images in visible spectrum, which are not suitable for night time detection. Infrared (IR) or thermal imaging is often adopted for night time due to its capability of capturing the emitted energy from pedestrians. The detection process firstly extracts candidate pedestrians from the captured IR image. Robust feature descriptors are formulated to represent those candidates. A binary classification of the extract features is then performed with trained classifier models. In this paper, an algorithm for pedestrian detection from IR image is proposed, where an adaptive fuzzy C-means clustering and convolutional neural networks are adopted. The adaptive fuzzy C-means clustering is used to segment the IR images and retrieve the candidate pedestrians. The candidate pedestrians are then pruned using human posture characteristics and the second central moments ellipse. The convolutional neural network is used to simultaneously learn relevant features and perform the binary classification. The performance of the proposed algorithm is compared with state-of-the-art algorithms on publicly available data set. A better detection accuracy with reduced computational accuracy is achieved.

1 Introduction and Related Work

In recent years with the increased interest in security and surveillance applications, advanced driver assistant systems, autonomous vehicle systems, and human behaviour analysis, pedestrian detection has become an important research problem in the computer vision research community. The accurate and robust detection of pedestrians is a challenging task. Some of the challenges in visible spectrum pedestrian detection include appearance variations, illumination variations, occlusions, human motion variations, and background noises. Additionally, owing to the sensor characteristics of the visible spectrum camera, the aforementioned challenges become more pronounced during the night time and bad weather conditions [1]. Consequently, researchers have sought to use infrared (IR) cameras for pedestrian detection. Unlike the visual camera, the IR camera captures the brightness intensity corresponding to the temperature and radiated heat of objects in the scene. Moreover, the IR camera is also independent of illumination variations, facilitating its use both during the day and night.

The IR image pedestrian detection algorithm, typically, consists of three components: candidate pedestrian detection, feature extraction, and feature classification. In IR images, instead of adopting a sliding window framework [2], researchers extract candidate pedestrians from the image to reduce the computational complexity. The candidate pedestrians are extracted using intensity information [3, 4], intensity gradient information [5], or motion information [6]. Bin et al. [5] utilised the vertical-horizontal intensity gradient information to extract candidate pedestrians. Alternatively, Bertozzi et al. [3] utilised the intensity information to identify and segment warm objects in the scene. Comparing the intensity-based segmentation approaches, we observe that the intensity gradient-based approaches are more susceptible to missing certain body parts, especially the limbs, in the extracted candidate pedestrians [5]. On the other hand, Dai et al. [6] utilised the pedestrian motion information within an EM framework to generate the candidate pedestrians. However, the motion-based segmentation requires a fixed camera, limited background motion, and cannot detect standing pedestrians. Given the extracted candidate pedestrians, researchers apply robust features descriptors to represent the candidate pedestrians. The feature descriptors should be able to discriminate the pedestrians from the background, while accounting for intra-pedestrian feature variations. The histogram of oriented gradients (HOG) and its variants are widely used for pedestrian detection in IR images [7, 5]. For example, Bin et al. [5] proposed a novel descriptor called the scattered difference of directional gradients (SDDG), while Liu et al. [8] proposed a pyramid entropy weighted HOG. The extracted features were then employed for binary classification using either template matching [9] or machine learning algorithms, such as the support vector machine (SVM) [7]. Examples of popular classification models include the AdaBoost [7], support vector machine (SVM) [7] and sparse representation classifiers (SRC) [2].

In this paper, we propose a pedestrian detection algorithm from IR image using an adaptive fuzzy C-means clustering and a convolutional neural network. The adaptive fuzzy C-means algorithm is employed to segment the IR images and retrieve the candidate pedestrians. Unlike the original fuzzy C-means algorithm, we adaptively estimate the required number of

clusters and fuse multiple clusters to retrieve the candidates. Utilising the human posture characteristics of walking uprightly, the second central moments ellipse is used to prune the set of candidate pedestrians. The pruned candidate pedestrians are then classified using a deep learning framework, i.e. the convolutional neural network. The convolutional neural network (CNN) functions as a joint feature extraction and feature classification model, eliminating the need for separately designing a robust feature descriptor and training a classification model. More importantly, the convolutional neural network has reported state-of-the-art performance on a variety of classification problems [10]; thus, motivated their use for IR image pedestrian detection in this paper. To validate the performance of the proposed algorithm, we experimented with the publicly available LSI data set, and compared with existing state-of-the-art algorithms [2]. The experimental results demonstrate the improved detection accuracy by the proposed algorithm.

The rest of this paper is organized as follows. In section 2, the proposed pedestrian detection algorithm is described. The experimental results are presented in section 3. Finally, this paper is summarized and concluded in section 4.

2 Pedestrian Detection Algorithm

The pedestrian detection is formulated as a binary classification problem, where a candidate pedestrian bounding box is classified as pedestrian or non-pedestrian. The proposed algorithm has three main components: candidate pedestrian detection using an adaptive fuzzy C-means clustering, candidate pruning using second central moments ellipse, and binary classification using trained convolutional neural network. An overview of the proposed algorithm is illustrated in Fig.1.

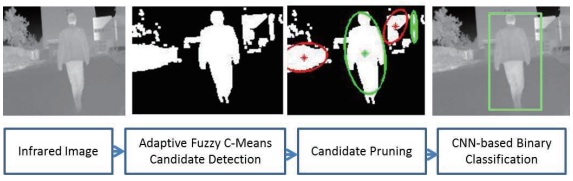

Figure 1. An overview of pedestrian detection framework.

Candidate Pedestrian Detection To reduce the computational complexity associated with the sliding window-based object detection, extraction of candidate pedestrians from the IR image as inputs to the classification model is suggested. In this paper, we adopted an adaptive fuzzy C-means clustering algorithm to extract the candidates in the IR image $I(a,b)$, where $x(a,b)$ represents the intensity of the pixel located at (a,b). The fuzzy C-means clustering is an iterative clustering algorithm [11] that soft partitions the input image into C clusters by minimizing an objective function J given below,

$$J(\mathbf{U},\mathbf{v}) = \sum_{c=1}^{C}\sum_{n=1}^{N}(u_n^c)^q d^2(x_n, v_c) \tag{1}$$

where $X = \{x_n\}_{n=1}^{N}$ is the set of N intensity levels in the IR image. C is the number of clusters, typically, manually specified. $\mathbf{U} = \{u_n^c\}$ is the fuzzy partition matrix, where u_n^c is degree of membership of n^{th} intensity in partition c. q is the membership weighting exponent and $\mathbf{v} = \{v_c\}_{c=1}^{C}$ is the weighted cluster centroid. $d^2()$ measures the distance between the intensity and cluster centroid. Qing et al. [11] utilised the histogram of image intensities, instead of raw image intensities, to minimise the computational complexity associated with obtaining the optimal $\mathbf{U}$ and $\mathbf{v}$. We direct the readers to [11] for a detailed explanation of the fuzzy C-means algorithm.

An important limitation of the fuzzy C-means algorithm [11] is the need to specify the number of clusters. In this paper, we adaptively estimate the cluster number using the image intensity information. More specifically, the number of clusters C is derived from the mean image intensity μ_I using the following measure, $C = \frac{\mu_I}{\eta}$ where η is an empirical constant. As shown in Fig. 2-b, the IR images acquired during the day time have a higher μ_I, and require more clusters than those acquired during the night or indoor scenarios (Fig.2-e). In our proposed algorithm, to estimate the candidate pedestrians, the optimised $\mathbf{U}$ is used to generate the c^{th} cluster map for the image $I(a,b)$ using the following assignment,

$$\mathbf{P}^c_{(a,b)} = u^c_{x(a,b)} \tag{2}$$

where $u^c_{x(a,b)} \in \mathbf{U}$ corresponds to the partition degree of membership for the pixel intensity $x(a,b)$. To robustly extract the candidate pedestrians, we identify $K \leq C$ cluster centroids $\{v_k\}_{k=1}^{K}$ in $\mathbf{v}$ with high intensities and adaptively fuse their corresponding cluster maps $\mathbf{P}^k$ to generate the candidate pedestrian map $\tilde{\mathbf{P}}$ by,

$$\tilde{\mathbf{P}} = \sum_{k=1}^{K} e^{-(k-1)}\mathbf{P}^k \tag{3}$$

where $k = 1$ corresponds to the highest intensity. By binarizing $\tilde{\mathbf{P}}$ and generating the connected components, the candidate pedestrians in the IR image are estimated. The proposed adaptive framework enables a robust extraction of candidate pedestrians for scenes with varying illumination as shown in Fig.2. In Fig.2-(b,e), where C is adaptively estimated as 4 and 2 respectively with η set to 25 and K set to 2, we can visualise the good segmentation results. However, in Fig.2-(c,f), where C is manually fixed as 3, we can observe the inferior segmentation results.

Pruning Candidate Pedestrians As humans tend to either stand or walk uprightly, we can utilise this characteristic to prune the set of extracted candidate pedestrians. We achieve this by utilising the spatial information of the candidate pedestrians connected components, in the binarized $\tilde{\mathbf{P}}$. More specifically, for each blob we estimate their second central moments ellipse and derive their corresponding properties, i.e. the centroid, major axis length, and orientation between the major axis and the horizontal. The major axis length and orientation information are then

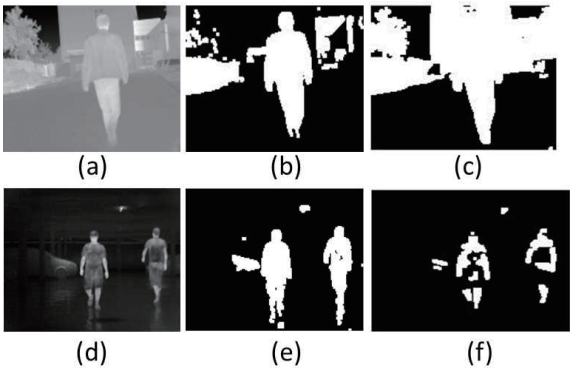

Figure 2. An example of (a-d) day and night time IR images, (b-e) adaptive segmentation results and (c-f) noisy segmentation results.

used to prune the candidate pedestrians using empirical thresholds. Finally, bounding boxes of multiple scales are extracted around the remaining candidate pedestrians, and given as input to the convolutional neural network. An illustration of the pruning is shown in Fig.3.

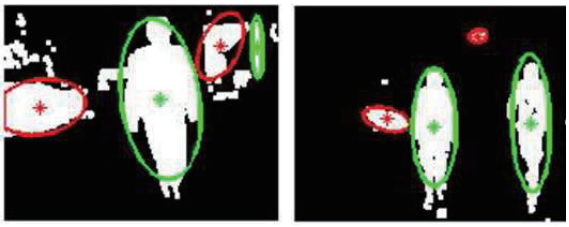

Figure 3. An illustration of candidate pedestrian pruning, where red ellipses represent the pruned candidates, and green ellipses represent the unpruned or selected candidates. Note we only illustrate the candidates with major axis lengths greater than the specified threshold.

Convolutional Neural Network In this paper, a binary classification of the candidate pedestrians was performed using a trained convolutional neural network. The convolutional neural network (CNN) can jointly learn the discriminative pedestrian features and the binary classification model [12]. However, to the best of our knowledge, they have not been used for pedestrian detection in IR images. CNN is based on the multilayer perceptron (MLP) and contains a multilayered architecture. The multi-layered architecture inherently contains a feature learning stage and a feature classification stage, each with multiple layers. The feature learning, or initial layers, in the CNN contains the convolutional layers. In the convolutional layers, learnt filters (weights) generate output feature maps by means of convolution with the input feature maps in the preceding layers, and the application of rectified linear units-based activation function. Additionally, to account for rotational, translation, scale and illumination variations, the feature learning stage contains the pooling layer and the local response normalisation. Pooling layers function as subsampling layers, and are used to summarize small neighbours or blocks in the convolutional layer. The pooling is done by calculating either the average or maximum value within a block. Multiple learnable filters in the various convolutional layers extract discriminative features, which are then fed to the fully connected networks in the deeper layers of the CNN. The final layer of the CNN contains the output neurons, which assign confidence scores to the pedestrian and non-pedestrian class.

CNN Architecture In the proposed algorithm, the CNN architecture is as follows. The input layer consists of IR image pixels with the size 128×128. Bounding boxes of different scales are resized before being given as the input to the CNN. The first layer consists of a convolutional layer with 96 filters of size 9×9 with relu activation function, maximum pooling with filter size 3×3 and a local response normalisation over a 5×5 window. The second layer consists of a convolutional layer with 256 filters of size 5×5 with relu activation function, maximum pooling with filter size 3×3 and a local response normalisation over a 5×5 window. The third and fourth layer each contains a convolutional layer with 384 filters of size 3×3 with relu activation function. The fifth layer consists of a convolutional layer with 256 filters of size 3×3 with relu activation function and a maximum pooling with filter size 3×3. The sixth and seventh layer consists of the fully connected network with 4096 neurons with relu activation function and dropout ratio of 0.5. The output layer consists of two output neurons, with softmax function, providing a score for each class. Using the class scores, each candidate bounding box is classified as either pedestrian or non-pedestrian, by assigning the label of the class with the highest score.

3 Experimental Results

Data Set and Parameter Setting The proposed algorithm was evaluated with the LSI public data set, which contains IR sequences acquired both during the day and night. In the experimental section, we performed a comparative analysis with the baseline algorithm [2]. Moreover, we also evaluated our proposed CNN approach with the sliding window-based CNN approach. The test sequences contains 2084 IR images with ground truth bounding boxes while the training data contains 8000 pedestrian or positive samples and 8000 negative samples. The algorithm was implemented on Linux using the GPU-based Caffe, deep learning framework [12].

Evaluation We evaluated our algorithm using the PASCAL measure [1], given in Eqn (4), after performing non-maximal suppression for merging multi-scale detections. A detected bounding box (BB_{dt}) is considered for potential match, if they overlap sufficiently (50% or 0.5) with the ground truth bounding box (BB_{gt}). To compare the performances, a log-log plot is adopted to present the miss rate against the false positives per image (FPPI). Additionally, the miss rate at nine FPPI rates are averaged, and the log-average

miss rate is represented for each detector.

$$\frac{area(\mathrm{BB}_{dt} \cap \mathrm{BB}_{gt})}{area(\mathrm{BB}_{dt} \cup \mathrm{BB}_{gt})} > \text{overlap ratio} \qquad (4)$$

Results Bin et al. [2] presented a sliding window-based pedestrian detection framework, where they evaluated combinations of multiple feature descriptors (HOG, histogram of sparse codes and phase congruency) and feature classifiers (SVM and SRC), and reported the best detection accuracies with HOG-SRC combination. Thus, we consider the HOG-SRC, in addition to the widely used HOG-SVM IR pedestrian detection framework, as our baseline algorithms. To ensure fair comparison, we replaced the sliding window framework in the baseline algorithm with the proposed adaptive fuzzy C-means-based candidate pedestrian detection. The results given in Fig.4 show that the proposed algorithm has lower log average miss rate than the baseline algorithm. For completeness, the sliding window-based HOG-SVM reported a log average miss rate of 79%, while the sliding window-based HOG-SRC reported a log average miss rate of 55%, which are inferior to their corresponding modified versions.

To evaluate the computational complexity of the proposed algorithm, we comparatively evaluated the sliding window-based CNN with the candidate pedestrian-based CNN. In the experiments, the sliding window-based CNN reported an log-average miss rate of 38% while taking 20 sec per frame. On the other hand, the candidate pedestrian detection-based CNN reported an average log-average miss rate of 34% (Fig.4), while taking 2.5 sec per frame. Thus, based on the experimental results, the proposed algorithm demonstrates a better detection accuracy than the baseline algorithms. Additionally, we also report significant reduction in the computational complexity while using the candidate pedestrian detection framework.

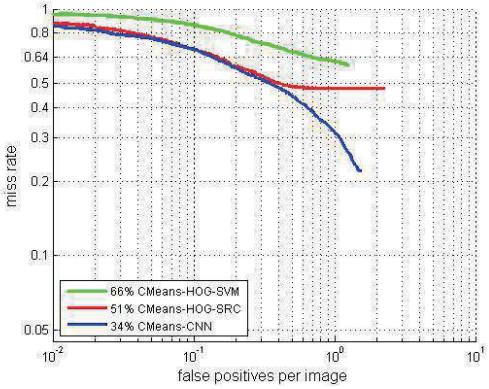

Figure 4. Evaluation of our proposed algorithm with the baseline algorithm.

4 Conclusion

In this paper, we propose an adaptive fuzzy C-means and convolutional neural network based pedestrian detection algorithm for IR images. The adaptive fuzzy C-means algorithm estimates candidate pedestrians in the IR image robustly for both day time and night time scenes. Additionally, we utilise the human upright posture characteristics to prune the candidate pedestrians using the second central moments ellipse. Finally, the convolutional neural network is used for binary classification. The experimental results show that the proposed algorithm has a better performance that the baseline algorithm in terms of detection accuracy. Moreover, compared with the sliding window framework, the candidate pedestrian detection framework significantly reduces the computational complexity. In future work, we will evaluate on algorithm on larger data sets, and incorporate tracking to further improve the detection accuracy and computational complexity.

References

[1] 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.

[2] B. Qi, V.John, Z.Liu, and S.Mita. Use of Sparse Representation for Pedestrian Detection in Thermal Images. In *IEEE Computer Vision and Pattern Recognition Workshop:Perception Beyond Visible Spectrum*, 2014.

[3] M. Bertozzi, A.Broggi, M.Carletti, A.Fascioli, T.Graf, P. Grisleri, and M. Meinecke. IR Pedestrian Detection for Advanced Driver Assistance Systems. *Lecture Notes in Computer Science*, 2781:582–590, 2003.

[4] S. Liu, Yupinluo, and S. Yang. Shape-based Pedestrian Detection in Infrared Images. *Journal of Information Science and Engineering*, 23:271–283, 2007.

[5] B.Qi, V. John, Z.Liu, and S.Mita. Pedestrian Detection from Thermal Images with A Scattered Difference of Directional Gradients Feature Descriptor. In *IEEE Intelligent Transportation Systems Conference*, 2014.

[6] C. Dai, Y. Zheng, and X. Li. Pedestrian detection and tracking in infrared imagery using shape and appearance. *Computer Vision and Image Understanding*, (2-3):288–299, 2007.

[7] L. Zhang, B. Wu, and R. Nevatia. Pedestrian detection in infrared images based on local shape features. In *IEEE Computer Vision and Pattern Recognition*, 2007.

[8] Q. Liu, J. Zhuang, and J. Ma. Robust and fast pedestrian detection method for far-infrared automotive driving assistance systems. *Infrared Physics & Technology*, 60:288–299, 2013.

[9] Y. Fang, K. Yamada, Y. Ninomiya, B. Horn, and I. Masaki. A shape-independent method for pedestrian detection with far-infrared images. *IEEE Transactions on Vehicular Technology*, 53:1679–1697, 2004.

[10] A.Krizhevsky, I. Sutskever, and G.Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[11] Y.Qing, H.Hua, and X. Qiang. Histogram based fuzzy c-mean algorithm for image segmentation. In *IAPR International Conference on Pattern Recognition*, pages 704–707, 1992.

[12] 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 arXiv:1408.5093*, 2014.