Infrared Physics and Technology 94 (2018) 69–77

Contents lists available at ScienceDirect

Infrared Physics & Technology

journal homepage: www.elsevier.com/locate/infrared

Regular article

Multispectral pedestrian detection based on deep convolutional neural networks

Ya-Li Hou*, Yaoyao Song, Xiaoli Hao, Yan Shen, Manyi Qian, Houjin Chen

School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing, China

ARTICLE INFO

Keywords:
Pedestrian detection
Multispectral information
CNNs

ABSTRACT

Vision-based pedestrian detection that can last all day is crucial in advanced driver-assistance systems (ADAS), autonomous vehicles and video surveillance. Based on the fact that the wavelength of human body radiation falls around 9.3 μm, thermal images have distinctive advantages for pedestrian detection in the nighttime. With the recent success of convolutional neural networks (CNNs) in the vision community, how to properly explore information in color and thermal images using CNNs-based methods attracts the attention of researchers. Previous CNN-based multispectral pedestrian detectors focus on the design of network architectures. The main contributions of this paper are as follows. First, pixel-level image fusion are also extensively evaluated based on a state-of-the-art CNNs-based framework in the daytime and nighttime for pedestrian detection. Second, the effective strategies to combine pixel-level fusion methods and CNN-fusion architectures are studied based on extensive experimental results. Two combination strategies are designed and an exhaustive experimental analysis is performed to evaluate different combinations for all-day pedestrian detection. Extensive results based on a multispectral pedestrian benchmark show that some pixel-level image fusion methods can achieve similar or even better performance than CNN-fusion architectures, which emphasizes the importance of pixel-level fusion in CNN-based pedestrian detectors. The combination of both can usually more properly exploit multispectral information and further boost detection performance.

1. Introduction

Vision-based pedestrian detection is a crucial but challenging problem in many autonomous systems. In [1,2], advances before 2014 have been intensively reviewed and investigated. Most work focused on the detection of pedestrians in visible-spectrum images. However, methods based on visible images usually work poorly in the nighttime. Since ambient lighting has less of an effect on thermal imaging, long-wavelength infrared (thermal) cameras are widely used to improve detection performance in the nighttime [3]. In addition, with the recent decrease in prices, thermal cameras have become attractive for more and more commercial applications. In advanced driver assistance systems (ADAS) [4,5], unmanned aerial vehicles (UAVs) [6] and surveillance systems, thermal images have attracted people's great attention.

Effectively exploring the clues provided by both visible and thermal images is an important topic for multispectral pedestrian detection. Recently, Jin et al. [7] has given a survey of visible and infrared image fusion methods, in which applications of the fusion methods, method analysis and quality measures are reviewed. Here, we mainly focus on related works for multispectral pedestrian detection. Usually, the fusion of information from visible and thermal cameras can be divided into three levels: pixel-level, feature-level and decision-level fusion. In pixel-level methods, pixels are determined from a set of image pixels or other forms of image parameters at the lowest physical level. There is a great number of pixel-level fusion methods. Laplacian pyramid (LP) fusion [8,9], fusion based on wavelet transform [10–12], fusion-based on curvelet transform [13–15] are several typical methods. [16] is a hybrid method, in which discrete stationary wavelet transform (DSWT), discrete cosine transform (DCT) and local spatial frequency (LSF) are combined for the image fusion. In [15], wavelet and curvelet transform were used to fuse the images for better pedestrian detection. In 2010, Choi and Park [17] developed a variant of joint bilateral filter to facilitate later human detection. In feature-level methods, information from multimodal images is combined in the form of image feature descriptors [18,19]. In 2015, Hwang et al.[20] proposed a multispectral pedestrian detector that fused a variety of features from visible and thermal images in aggregated channel features. In [21], an exhaustive experimental analysis was performed to demonstrate the advantages of multispectral detection methods based on different state-of-the-art features and classifier combinations. The researches of decision level

* Corresponding author.
E-mail address: ylhou@bjtu.edu.cn (Y.-L. Hou).

https://doi.org/10.1016/j.infrared.2018.08.029
Received 25 December 2017; Received in revised form 30 August 2018; Accepted 30 August 2018
Available online 01 September 2018

image fusion is the least, and mainly are applied in face recognition [22,23]. Torresan et al. [24] is an example of the decision-level fusion methods for pedestrian detection, in which the detection and tracking results of visible and thermal images are merged.

With the recent success of convolutional neural networks (CNNs) in the vision community, the question of how to properly explore multimodal information using CNNs-based methods has attracted the attention of researchers [25–27]. Some multispectral pedestrian detectors based on CNNs have also appeared recently [28,29]. In [28], two CNN-fusion architectures were investigated based on the R-CNN detection framework. The results show that the late-fusion CNN significantly outperforms the early-fusion architecture. In [29], four fusion architectures that integrate two-branch CNNs at different stages were tested based on the Faster R-CNN framework. Among the four architectures, Early, Halfway and Late Fusion perform feature-level fusion. The Score Fusion combines the decision results of color and thermal branches and is a decision-level fusion. The Halfway Fusion model had the best performance in the evaluations.

Although CNN-based methods has achieved great success in pedestrian detection, fusion of multimodal images in CNN-based detectors have not been well studied. It should be noted that the above work focused on fusion based on CNNs, the importance of pixel-level image fusion was not considered. On the other hand, [20] has recently provided a large multispectral pedestrian dataset with well-aligned color-thermal image pairs, which makes pixel-level image fusion feasible. Hence, This paper aims to fill this gap and further explore the "optimal" way to fuse visible and thermal image channels in CNN-based multispectral pedestrian detectors. In this paper, we try to answer two questions: **(1) How does the performance of image-fusion methods compare with other CNN-fusion architectures? (2) Are the fusion methods complementary? Is it necessary to fuse the images before inputting them into CNN-fusion architectures?**

2. Methods

Since Girshick et al. developed the R-CNN [30], Fast R-CNN [31] and Faster R-CNN [32] frameworks, CNNs have become a popular and effective tool for object detection and recognition. To find a proper way to fuse visible and thermal images in CNN-based detectors, the performance of different fusion methods will be investigated in this paper. The Single Shot Detector (SSD) framework [33] is used as a baseline system in our evaluations. SSD is a single-stage object-detection method that unified four steps—object proposal, feature extraction, classification and regression—into a single network that can achieve comparable accuracy to Faster R-CNN, but much faster. We use 300 × 300 SSD to reduce the training time, in which the input images are resized to 300 × 300 before injecting into the networks.

2.1. Pixel-level image fusion

As we know, thermal images are usually intensity images, to fuse thermal images (T) with tri-chromatic visible images at the pixel level, the flow chart in Fig. 1 is employed. Visible images are first converted to HSI/YUV color space and image fusion is performed between the I/Y component and the thermal component. The fused RGB images are then reconstructed using the fused I/Y component and the two remaining components. The fusion scheme can preserve color information better than fusing thermal images with each R, G and B component separately. Finally, the fused images are injected into the SSD base network. In this paper, three classic transformation-based methods and one spatial-based fusion method are tested, including LP fusion [8], wavelet fusion [10], curvelet fusion [13] and fusion based on a joint bilateral filter.

The transformation-based methods usually include three main stages: transformation, fusion and reconstruction. In the current evaluations, the parameters in each method are empirically chosen such that no obvious artifacts can be observed in the fused images. More deliberate parameter settings and fusion rule selection will be studied in the future. In the LP method, a five-layer pyramid is established and the Gaussian kernel size is 3 × 3. To fuse multimodal images, the low-frequency component with the highest average gradients within the 3 × 3 neighborhood window is used to preserve details in the images and the high-frequency components with the maximum coefficients are used. In the wavelet fusion method, two-layer wavelet decompositions are performed for both the thermal images and the I components. The Daubechies wavelet "db3" is used. The average of the low-frequency components and the maximum high-frequency components are used for the fused image. In the curvelet fusion method, two scales are used; One orientation is used for the first scale and 16 orientations are used for the second scale. To fuse the images, the coefficients of the first scale are averaged to merge the holistic information of both modalities, and the maximum coefficients of the second scale are used to preserve edges and textures.

In [17], a spatial-based image-fusion method is specifically designed for human detection and based on a variant of the joint bilateral filter. A new joint bilateral filter is developed to fuse a visible image and a segmented thermal image. Segmentation of the thermal image is applied to remove the effects of background regions and human shadows and to facilitate later human detection based on differences between the images. We did not use segmentation in our evaluations. The reasons are as follows: (1) Threshold selection is always a big headache for the binarization of the grayscale thermal image. (2) Instead of removing the background regions, the background can be used as useful clues for pedestrian detection in CNN-based algorithms. Hence, different from [17], the original joint bilateral filter is used in our tests to fuse multimodal images. The equation to fuse the images based on the joint bilateral filter is shown in (1).

$$F_p = \frac{1}{k(p)} \sum_{p' \in \Omega} g_d(p' - p) g_r\left(A_{2p} - A_{2p'}\right) A_{1p'} \tag{1}$$

In (1), F_p is the intensity of point p in the fused image. $A_{1p'}$ is the intensity of point p' in the base image, and A_{2p} is the intensity at p in the other image. We follow [17] and use the Y component to fuse with the thermal image. It is usually believed that visible images are important for daytime pedestrian detection, and thermal images play an important role in nighttime detectors. Hence, to preserve most of the information in visible images, in the daytime, the Y component of visible images is used as the base image A_1 and the thermal image is used as the other image A_2. In the nighttime, the thermal component is used as A_1 and the Y component is used as A_2. In a joint bilateral filter, the weights consist of two functions: a spatial domain function g_d and an edge-stopping function g_r. The function g_d is determined by the distance between the pixels, whereas the function g_r is determined by the intensity difference in A_2. Typically, these functions are Gaussian-shaped and their standard deviation parameters are σ_d and σ_r, respectively. Ω is a neighborhood window region around the pixel p, and k(p) is a normalization factor. In our evaluations, σ_d and σ_r are both 6 and Ω is a 3 × 3 local region.

Finally, the fused images are converted to RGB and injected into an SSD detector. Due to limited space, sample images of fusion results at day and night are provided in Appendix A. (Figs. A1 and A2 show some example image fusion results on KAIST reasonable day and night subsets respectively. It seems that wavelet-fusion, Laplacian Pyramid fusion and curvelet-fusion obtain similar fusion results and preserve most visible information. Results based on the joint bilateral filter look more like thermal images in the nighttime.) For the SSD detector, the weights for the original VGG16 layers [34] in the base network are pre-trained on the ILSVRC CLS-LOC dataset [35], and the weights for all the other convolutional layers are initialized using the "Xavier" method [36]. The original VGG16 model has 13 convolutional layers and 3 fully-connected layers and small twists are performed to be used as the base network of the SSD detector [33]. To achieve better detection