Deep Background Subtraction of Thermal and Visible Imagery for Pedestrian Detection in Videos

Yijun Yan[1], Huimin Zhao[2,3], Fu-Jen Kao[4], Valentin Masero Vargas[5], Sophia Zhao[1], Jinchang Ren[1]

[1] Dept. of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK
[2] School of Computer Science, Guangdong Polytechnic University, Guangzhou, China
[3] The Guangzhou Key Laboratory of Digital Content Processing and Security Technologies, Guangzhou, China
[4] Institute of Biophotonics, National Yang-Ming University, Taipei, ROC
[5]Department of Computer Systems and Telematics Engineering
Universidad de Extremadura, Badajoz, Spain

Abstract. In this paper, we introduce an efficient framework to subtract the background from both visible and thermal imagery for pedestrians' detection in the urban scene. We use a deep neural network (DNN) to train the background subtraction model. For the training of the DNN, we first generate an initial background map and then employ randomly 5% video frames, background map, and manually segmented ground truth. Then we apply a cognition-based post-processing to further smooth the foreground detection result. We evaluate our method against our previous work and 11 recently widely cited method on three challenge video series selected from a publicly available color-thermal benchmark dataset OCTBVS. Promising results have been shown that the proposed DNN-based approach can successfully detect the pedestrians with good shape in most scenes regardless of illuminate changes and occlusion problem.

Keywords: Deep neural network (DNN), Video salient objects, Pedestrian detection.

1 Introduction

Background subtraction is always a crucial step for pedestrian detection. For outdoor surveillance in the urban setting, there are a tremendous amount of available video data. However, most data such as background scenery are redundant, only pedestrians are meaningful information and are necessary to extract. Meanwhile, if the background scenery can be get rid of, both storage and computing resources will be saved. For pedestrian detection, visible camera and thermal imagery are two popularly used sources of image modalities, though not necessarily in a combined solution [5]. However, either visible image or thermal image has their advantages and disadvantages. The visible image can show detailed color information, but the two biggest challenge for visible images are hard shadow and illumination change. Although many back-

ground subtraction methods have been explored in recent years, the performance is still not good enough when facing these challenges. Since the object is detected by its temperature and radiated heat, thermal image can eliminate the influence of color and illumination changes on the objects' appearance [10] in any weather conditions and at both day and night time. However, in some cases e.g. occlusions, the thermal camera may fail to detect the object properly. Moreover, it will detect any objects with surface temperature. Therefore, in this paper, we present a deep pedestrian detection model that fuses the information from both visible and thermal images.

Deep learning technique is a very hot topic recently and has been widely explored in many areas [7; 20; 25]. Fusion strategy is also very popular in many researches (e.g. image retrieval[21; 22], image recognition[4], ROI detection[14; 15; 27], saliency detection[23], etc) since it can help break the limitation of the single feature. In our work, we first generate the background map for both visible and thermal video sequences. Then, we extract fused features of both visible and thermal image by a DNN to predict the foreground map for pedestrian detection. To prepare the training data patches for DNN, we select 5% video frames and over-segment each video frame, corresponding ground truth and background map into a large number of subsets of the scene. Finally, a cognition-based refinement is applied to smooth the fore-ground map from DNN.

The outline of this paper is as follows: Section 2 illustrates the framework of the pro-

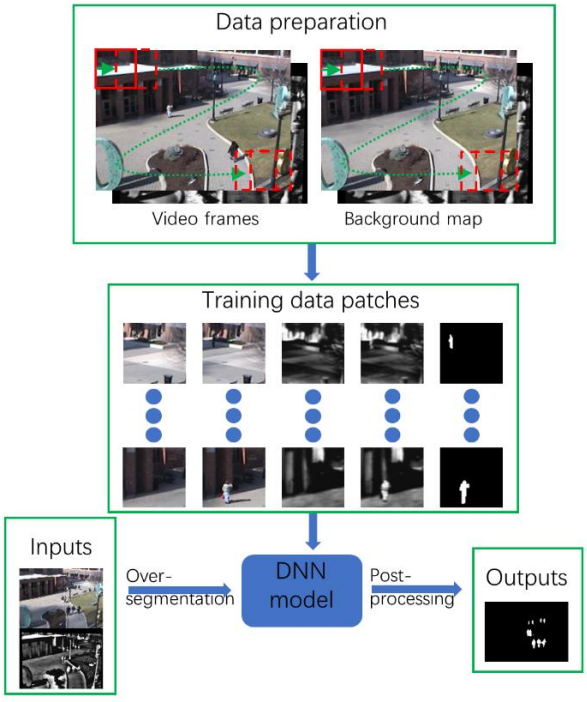

Fig. 1. The framework of our proposed DNN-based background subtraction method.

posed method. Section 3 describes the detail of our DNN model. Experimental results are presented and discussed in Section 4. Finally, some concluding remarks and future work are summarized in Section 5.

2 Overview of the proposed method

In this paper, we proposed a DNN-based background subtraction model for foreground detection of pedestrians on both visible and thermal video sequences. In the first step, we generate an initial background map for visible and thermal video frames, respectively. Then we over-segment each frame into plenty of regions which overlap with each other. To achieve that, we apply a moving window of size 64*64 on each frame (see 'Data preparation' in Fig. 1). Then the train image patches can be obtained in which the information of visible frames, visible background, thermal frames, thermal background and ground truth are included (see 'Training data patches' in Fig. 1). After the deep neural network is trained, we input both visible and thermal frame to the network and get the foreground detection result of pedestrians.

3 DNN-based foreground detection

We train our proposed DNN with pairs of visible and thermal image patches from video and their background frames. In this section, the data preparation step for network training and network architectures will be explained in detail.

3.1 Data preparation

To train the DNN, our training data patches consist of the video frame, background map and ground truth. In order to generate background map for visible and thermal image, we compute a median map of N frames randomly selected from the video sequence. Although this pixel-wise temporal median filtering doesn't work very well if there are a lot of moving objects in the scene[1], it works very well in our selected video frames. Because, pedestrians don't move very fast in these frames, and the number of pedestrians is not too much. The other advantage of this method is its simplicity. It doesn't need too much computation cost.

To generate the image patches of visible and thermal video frames, we first random select 5% data samples from the video sequences, which contains various challenging video scenes and their ground truth segmentation. Then we segment each frame into plenty of subsets of a scene with size 64*64 and scale the pixel value to [0,1]. Therefore, our inputs of the network have a size of 64*64*8 which includes the visible frame, thermal frame, visible background and thermal background. Our outputs of the network have a size of 64*64*2 since we regard the foreground detection as a binary classification problem. An example of our training patches is shown in Fig. 2. Last but not least, we perform a mean subtraction on each pixel as suggested in many works[8; 12].

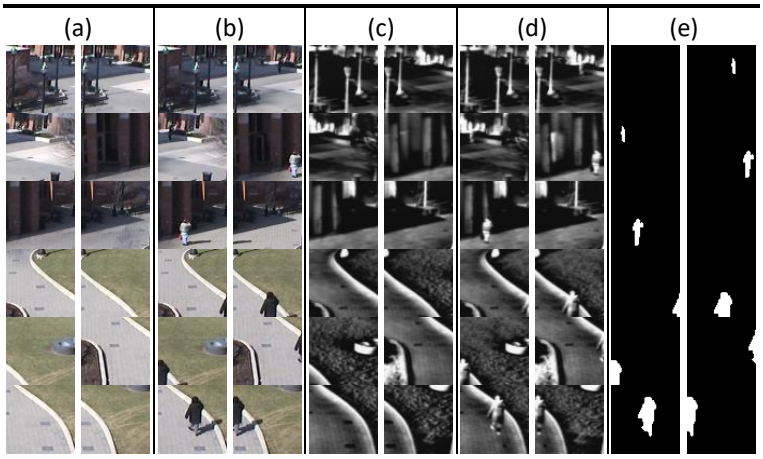

Fig. 2. Visualization of training image patches: (a) visible background patches, (b) visible input patches from visible video sequences, (c) thermal background patches, (d) thermal input patches from thermal video sequences. (e) ground truth patches.

3.2 Network architecture and training

The architecture of the proposed deep neural network is shown in Fig. 3. The network contains 4 convolutional layers where several convolution kernels are used to extract different features. For each layer, the size of the kernel is 3*3 with a stride of 1 and pad of 1. The number of output of the layer $L \in [1,2,3,4]$ is [48,246,512,2], respectively. The activation function we use is the Rectified Linear Unit (ReLU). It can improve the representation ability of the network model. In addition, the batch normalization (BN) is also applied for first two layers. The advantage of batch normalization layer in DNN is it makes us care less about initialization and can use much higher learning rate. We train our DNN model for 100 epochs with a mini batch size of 20 image patches, and a learning rate of $\alpha = 10^{-3}$. For the loss function, we use the softmax loss, which is defined as follows:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})] \quad (1)$$

where, N is the number of training image patches, y_i is the ground truth, $\hat{y_i}$ is the predicted foreground result.

3.3 Post processing

In order to generate the final foreground map and make the result close to human perception, we put a shape constrained morphological refinement to the results from the DNN model. Here, we define a function $D(\cdot)$ that can dilate all the potential objects with a shape based structuring element. The width and height of the rectangle are

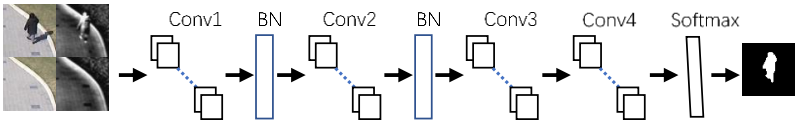

Fig. 3. Architecture of proposed DNN for background subtraction.

defined as $2n+1$ and $2n+3$ ($n \in Z_0^+$), and we set n=0 because we just want to smooth the edge for each object and connect the small gap between some object pieces. By doing so, the shape of the object will have continuity, which matches human perceptions.

4 Experimental results

4.1 Dataset description and evaluation criteria

To validate the effectiveness of deep-neural-network-based foreground detection method, we tested our method on three challenging visible and thermal video sequence pairs from a publicly available database '03OSU Color-Thermal Database'. These data are recorded on Ohio State University campus at different times-of-day with different camera gain and level settings. Thermal sequences are captured by Raytheon PalmIR 250D thermal sensor and color sequence are captured by Sony TRV87 Handycam color sensor. All the frames in both sequences have a spatial resolution of 320*240 pixels. The number of frames in each video sequence is Sequence-1:2107, Sequence-2:1201, Sequence-3:3399, respectively. Fig. 4 shows some examples of the visible and thermal video sequences. It can be seen that video sequences contain several people, some in groups, moving through the scene, and regions of dark shadows and illumination changes in the background.

In our experiment, we do both qualitative and quantitative assessment on some manually segmented silhouettes and we benchmark with our previous work [24] and other 11 methods which have been published in recent years and widely cited i.e. GMG[6], IMBS[3], LOBSTER[16], MultiCue[13], SuBSENSE[17], T2FMRF [26], ViBe[2], FA-SOM[11], PBAS[9], DECOLOR[28] and our previous work[24]. Three commonly used criteria i.e. precision, recall and F-measure, are adopted in our experiments to quantitatively assess the performance of proposed foreground detection method and other benchmarking methods. Three criteria are listed in the following:

- Precision (P): $$P = \frac{T_p}{T_p + F_p} \quad (2)$$
- Recall (R): $$R = \frac{T_p}{T_p + F_n} \quad (3)$$
- F-measure (FM): $$FM = \frac{2 \cdot P \cdot R}{P + R} \quad (4)$$

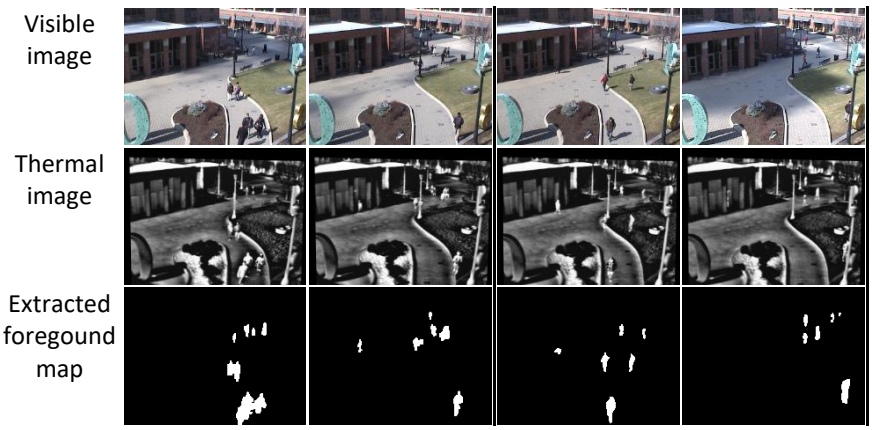

Fig. 4. Visible images, thermal images and results of our foreground detection method.

where T_p, F_p, and F_n respectively refer to the number of correctly detected foreground pixels of the pedestrians, incorrectly detected foreground pixels (false alarms), and incorrectly detected background pixels (or missing pixels from the object). Specifically, these three numbers can be calculated by comparing the binary masks of the detected image and the ground truth. Furthermore, since the database doesn't have ground truth, we obtain a manual segmentation of the pedestrian regions in plenty of frames from selected video sequences.

4.2 Assessment of foreground detection method

To better evaluate the quality of our foreground map and the results from other methods, we do the quantitative comparison in terms of precision, recall and F-measure in Table 1, and qualitative comparison in Fig. 5. Since most benchmarking background subtraction methods are designed for visible images, however, the selected video sequences contain both visible and thermal imagery. Therefore, for a fair comparison, we employed a fusion strategy from our previous work [24] for each method. The final foreground map is integrated by both visible foreground map and thermal foreground map which are generated by those methods on visible and thermal images respectively. In Table 1, it can be seen that our DNN-based method outperforms other mainstream background subtraction approaches on selected three challenge video sequences in terms of precision and F-measure. Although MultiCue[13] yield slightly higher recall (0.2%) than proposed method, its precision is the lowest among all the methods which mean its foreground map contains too much false alarm. For qualitative evaluation, we only show the results of our method and five benchmarking methods. The results of other six methods (i.e. IMBS[3], MultiCue[13], T2FMRF [26], FA-SOM[11], PBAS[9], DECOLOR[28]) are not in this figure, due to the page limit and their relatively low F-measure value. As can be seen in Fig. 5, these methods can detect the pedestrians within close or middle range but not long range from the camera. In addition, affected by light change and weather condition, some details have

Table 1. Comparison of Precision, Recall and F-measure values.

Method	Precision	Recall	F-measure
GMG[6]	70.45	70.17	70.31
IMBS[3]	37.03	74.44	49.46
LOBSTER[16]	72.95	72.19	72.57
MultiCue[13]	26.02	88.78	40.25
SuBSENSE[17]	69.31	76.87	72.89
T2FMRF [26]	50.78	29.93	37.66
ViBe[2]	74.06	64.38	68.88
FA-SOM[11]	38.45	82.86	52.53
PBAS[9]	73.02	33.60	46.02
DECOLOR[28]	50.71	91.75	65.32
Previous work[24]	70.16	87.97	78.06
Proposed	**89.22**	**88.54**	**88.88**

been lost. For example, some pedestrians' shapes in GMG are fractured (e.g. the pedestrian group in 2nd and 3rd images); some pedestrians that far away from the camera can't be detected in both LOBSTER and SuBSENSE (e.g. the person at up-right in 4th image), and their detection of individual person is melt and stick with each other; most pedestrians' shapes in Vibe are not integrated. Hence, these methods have good quantitative results but their qualitative results don't fit human's cognition. Although the foreground detection of our previous work has been improved a lot, we still can't detect the precise edge of people. However, thanks to the DNN that allow our proposed method can detect the single pedestrian or pedestrian group with more accurate shape even under such a challenge scene with dark shadow and illumination change.
In general, our proposed method yields the best performance in terms of quantitative and qualitative performance, but there are still rooms for further improvements. Our current DNN model doesn't contain too many layers which means the feature extraction of the visible and thermal image is not enough. In addition, due to the difference between visible and thermal imagery, each of them should have a typical DNN model. And the final foreground detection can be fused by a statistical fusion strategy.

5 Conclusion

In this paper, we proposed a background subtraction method using deep neural network for pedestrian detection in visible and thermal imagery. We first generate an initial background map by random median stage and build a lot of ground truth map

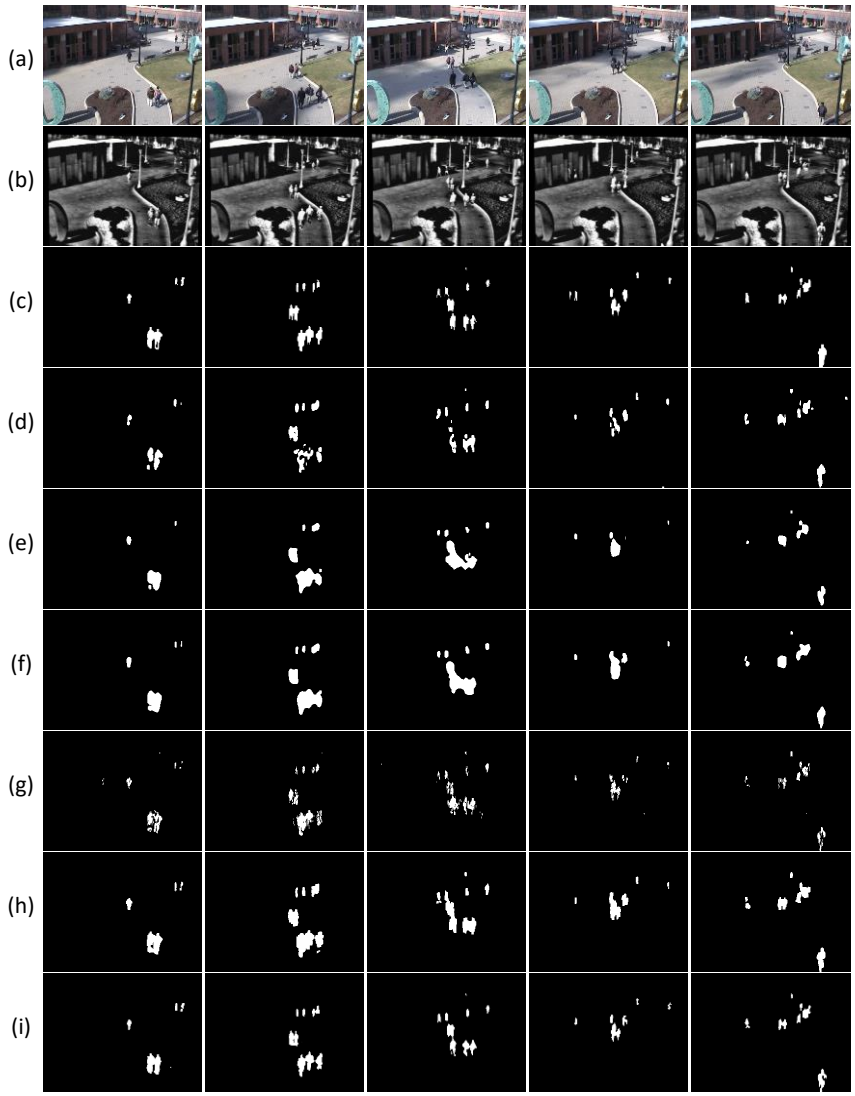

Fig. 5. Visual comparison, (a) RGB images, (b) thermal image, (c) ground truth, (d) GMG[6], (e) LOBSTER[16], (f) SuBSENSE[17], (g) Vibe[2], (h), previous work[24], (i) Proposed method.

by manual segmentation. After that, we put all data, initial background map and ground truth to train the DNN model. Finally, a shape constrained morphological refinement is applied to further improve the quality of the foreground detection result. The experimental results show that our approach yields better performance than other classical background modeling approaches in urban scenes. In the future, we would design two typical DNN model for visible and thermal imagery respectively, improve

the applicability of our background method approach in different types of scenarios and test our method on more challenge database such as CDnet[19] and BMC[18].

References

1. Babaee M, Dinh DT, Rigoll G (2017) A deep convolutional neural network for background subtraction arXiv preprint arXiv:170201731
2. Barnich O, Van Droogenbroeck M ViBe: a powerful random technique to estimate the background in video sequences. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, 2009. IEEE, pp 945-948
3. Bloisi D, Iocchi L Independent multimodal background subtraction. In: CompIMAGE, 2012. pp 39-44
4. Chai Y, Ren J, Zhao H, Li Y, Ren J, Murray P (2016) Hierarchical and multi-featured fusion for effective gait recognition under variable scenarios Pattern Analysis and Applications 19:905-917
5. Davis JW, Keck MA A two-stage template approach to person detection in thermal imagery. In: null, 2005. IEEE, pp 364-369
6. Godbehere AB, Matsukawa A, Goldberg K Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation. In: American Control Conference (ACC), 2012, 2012. IEEE, pp 4305-4312
7. Han J, Zhang D, Hu X, Guo L, Ren J, Wu F (2015) Background prior-based salient object detection via deep reconstruction residual IEEE Transactions on Circuits and Systems for Video Technology 25:1309-1321
8. He K, Zhang X, Ren S, Sun J Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. pp 770-778
9. Hofmann M, Tiefenbacher P, Rigoll G Background segmentation with feedback: The pixel-based adaptive segmenter. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, 2012. IEEE, pp 38-43
10. Kim D-E, Kwon D-S Pedestrian detection and tracking in thermal images using shape features. In: Ubiquitous Robots and Ambient Intelligence (URAI), 2015 12th International Conference on, 2015. IEEE, pp 22-25
11. Maddalena L, Petrosino A (2010) A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection Neural Computing and Applications 19:179-186
12. Nguyen TP, Pham CC, Ha SV-U, Jeon JW (2018) Change Detection by Training a Triplet Network for Motion Feature Extraction IEEE Transactions on Circuits and Systems for Video Technology
13. Noh S, Jeon M A new framework for background subtraction using multiple cues. In: Asian Conference on Computer Vision, 2012. Springer, pp 493-506
14. Ren J, Han J, Dalla Mura M (2016) Special issue on multimodal data fusion for multidimensional signal processing Multidimensional Systems and Signal Processing 27:801-805
15. Ren J, Jiang J, Wang D, Ipson S (2010) Fusion of intensity and inter-component chromatic difference for effective and robust colour edge detection IET image processing 4:294-301
16. St-Charles P-L, Bilodeau G-A Improving background subtraction using local binary similarity patterns. In: Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, 2014. IEEE, pp 509-515

17. St-Charles P-L, Bilodeau G-A, Bergevin R Flexible background subtraction with self-balanced local sensitivity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014. pp 408-413
18. Vacavant A, Chateau T, Wilhelm A, Lequievre L A benchmark dataset for foreground/background extraction. In: ACCV 2012, Workshop: Background Models Challenge, 2012.
19. Wang Y, Jodoin P-M, Porikli F, Konrad J, Benezeth Y, Ishwar P CDnet 2014: An expanded change detection benchmark dataset. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on, 2014. IEEE, pp 393-400
20. Wang Z, Ren J, Zhang D, Sun M, Jiang J (2018) A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos Neurocomputing 287:68-83
21. Yan Y, Ren J, Li Y, Windmill J, Ijomah W Fusion of dominant colour and spatial layout features for effective image retrieval of coloured logos and trademarks. In: Multimedia Big Data (BigMM), 2015 IEEE International Conference on, 2015. IEEE, pp 306-311
22. Yan Y, Ren J, Li Y, Windmill JF, Ijomah W, Chao K-M (2016) Adaptive fusion of color and spatial features for noise-robust retrieval of colored logo and trademark images Multidimensional Systems and Signal Processing 27:945-968
23. Yan Y et al. (2018) Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement Pattern Recognition 79:65-78
24. Yan Y et al. (2018) Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos Cognitive Computation 10:94-104
25. Zabalza J et al. (2016) Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging Neurocomputing 185:1-10
26. Zhao Z, Bouwmans T, Zhang X, Fang Y (2012) A fuzzy background modeling approach for motion detection in dynamic backgrounds. In: Multimedia and signal processing. Springer, pp 177-185
27. Zheng J, Liu Y, Ren J, Zhu T, Yan Y, Yang H (2016) Fusion of block and keypoints based approaches for effective copy-move image forgery detection Multidimensional Systems and Signal Processing 27:989-1005
28. Zhou X, Yang C, Yu W (2013) Moving object detection by detecting contiguous outliers in the low-rank representation IEEE Transactions on Pattern Analysis and Machine Intelligence 35:597-610