EasyChair Preprint

№ 7559

Detection of Moving Objects by Background Subtraction for Foreground Detection

Mukarram Safaldin and Nizar Zaghden

March 13, 2022

Detection of Moving Objects by Background Subtraction for Foreground Detection

Mukaram Safaldin; Nizar Zaghden

National School of Electronics and Telecommunications of Sfax, University of Sfax, Tunisia

malqadiry@yahoo.com; Nizar.zaghden@gmail.com

ABSTRACT:

Background subtraction for foreground detection has been commonly applied for varying usages to identify objects in motion within a scene, such as that in video surveillance. In fact, significant publications were noted in the last decade within this area of background modelling. Despite the several surveys noted in the literature, none has offered a comprehensive review in this field. Therefore, this paper elaborates both conventional and recent approaches in light of background modelling. Initially, the approaches listed in the literature were classified in terms of mathematical models. Next, these models were analyzed based on challenging scenarios that they managed, the challenges and issues are then summarized. After that, an enhanced method is proposed, resulting from hybridizing the weight optimizations from CNN with a set of customized features derived by the Viola- Jones detector to enhance the overall network's performance. The initial findings show that the proposed method superior the state-of-the-art models in terms of accuracy, recall, precision, and F-measure.

Keywords: Background subtraction; foreground detection; moving objects; object detection.

1. INTRODUCTION:

Analyzing and comprehending video sequences has gained much popularity amongst researchers. Multiple applications within this research domain (i.e., video surveillance [1–3], optical motion capture [4], & multimedia usage [5]) require detection of moving objects in the scene as the initial step. The fundamental operation sought refers to segregation of moving objects called 'foreground' from static information known as 'background'. The typical process applied here refers to background subtraction elaborately detailed in [6–8]. One easy technique for modelling the background is acquiring a background image that excludes object in motion. The background in certain settings is to no avail and can be altered if critical condition emerges, e.g., object removal or introduction and changes in illumination in the scene. Therefore, a more adaptive and robust background representation model is needed. The two issues linked with background subtraction are salient motion detection [10] and change detection [9]. Changes are detected between two images. Hence, background subtraction occurs when (1) changes stem from objects in motion and (2) two images are present: background and current. Meanwhile, salient motion detection seeks semantic regions and filters out insignificant regions. The very concept of saliency detection stems from human visual system - a simple, pre-attentive, and easy process. Hence, detection of salient motion reflects a scenario of background subtraction.

Background subtraction has earned popularity due to its capability to detect foreground from video streams, i.e., automated video surveillance and Human-Machine Interaction (HMI) [11-12], content-based video coding [13], anomaly detection [14], people counting [15], background substitution [16], visual analysis of human activities [17-18], visual observation of animals [19-20], and target tracking [21-22]. Background subtraction adheres to the similar scheme for detecting an object in motion [23]: background initialization, background maintenance, and foreground detection. Several settings, including uncluttered scene, static camera, static background, and constant illumination, appear crucial to successfully execute background subtraction. Nonetheless, actual settings pose challenging scenarios that interrupt foreground detection in videos. Background subtraction robustness is influenced by multiple aspects, including bootstrapping, difference in illumination, low frame rate, shadows, night videos, dynamic background, intermittent object motion, camouflage, and cameras in motion (hand-held cameras/aerial vehicles/pan-tilt-zoom cameras/mobile devices), elaborated in detail in [24-25].
The reminders of this article are as follows: Section 2 analyzes the existing literature. Section 3 summarizes the challenges and issues, then Section 4 illustrated the proposed model with initial results. Finally, the article is concluded in Section 5.

2. RELATED WORKS

2.1 Video surveillance

Increased installation of CCTV and advanced camera infrastructures has led to the emergence of intelligent video surveillance systems for automatic specific monitoring. The main goal for installing the video surveillance system is to automatically interpret a scene after analyzing interactions and motions of objects to hinder unpleasant incidences. Essentially, video surveillance systems ascertain safety and security aspects. In particular, abnormal behavior detection has gained popularity in this research domain [26].

For instance, a monitoring system was developed to detect fall among elderly within the setting of a home [27]. The ellipse-shape system is tailored to fit the body of the subject, while tracking of the head position identified change in the posture of the subject. Another automated surveillance system was devised to detect burglary via motion and posture analyses with low-cost hardware (e.g., consumer's camera) [28]. In a system that converts results in 2D to 3D space; abnormal crowd motion detection technique was deployed for uncontrolled outdoor setting [29]. Notably, the cluster size, orientation, and location generated by crowd had been applied to estimate crowd behaviour. Mismatch between cluster motion and prediction reflects higher chances of abnormal events. Recently, a system was initiated to identify real-time suspicious behaviour installed at shopping malls [30]. The system detects suspicious motion, including loitering activities and unattended cash register. Another smart surveillance system was introduced to detect interpersonal crime (e.g., harassment, trespass, & violence) in public transport and places by identifying positional change speed of the human subject [31].
The video surveillance system is comprised of object detection, tracking, and recognition, as well as behavioral analysis [32]. In object detection, the region of interest (ROI) is determined (e.g.,

vehicles/human in motion). The detected objects are grouped into predefined categories during object recognition step. Moving objects are identified and categorized into ‘car’ or ‘human’ in accordance to their features; i.e., color, shape, and pattern. Next, tracking of objects and analysis of behaviour are executed to detect suspicious events. Detection and recognition of object are crucial in determining the overall surveillance system performance, mainly because the following steps are highly reliant on the outcomes of these initial steps.

Nevertheless, several shortcomings have been linked with video surveillance system, such as occluded objects in video and disruptive noise due to change in illumination. One should note that such setbacks are higher in outdoor setting [33] due to poor quality of videos as a result of illumination changes and objects frequently identified far from camera. On top of that, various algorithms formulated in controlled setting exert poor performance in actual condition [34]. Most techniques are suitable for day-time surveillance due to heavy reliance on lighting [35], while abnormal incidences, i.e., crime, commonly happen at night. Thus, it is crucial to overcome the stated drawbacks by developing suitable video surveillance systems meant for outdoor setting.

2.2 Background subtraction

One viable method deployed in computer vision systems refers to background subtraction, which was initially applied for detecting objects in motion across video streams. Algorithms of background subtraction differentiates moving objects or better known as ‘foreground’ from the sequence background in the video stream without object details [36]. The vastly explored background subtraction method for video surveillance analysis since the 1990s is meant to identify moving objects from background prior to other intricate detection processes, i.e., people counting, invasion, and tracing [37].

The three stages in background subtraction algorithms are [38]: (1) Background initialization: development of a background model with certain number of frames and designed in multiple ways, i.e., neural network, statistical, and fuzzy. (2) Foreground detection: background model and present frame are compared. Connection is made to determine scene foreground via subtraction. (3) Background maintenance: images trained during the initialization step meant to update the background model are analyzed during the process of detection.

An object in motion for an extended time should be amalgamated into the background. Algorithms for background subtraction are grouped based on the technique of background model development, such as basic, cluster, statistical, fuzzy, and neuro-fuzzy modelling types. Upon computing the model in basic modelling, the variance between background image and present frame is modified in accordance with threshold. If the computed outcome exceeds that of threshold, the present frame pixels are segmented into foreground. Prior frame or static image functions as background in basic modelling. In fact, a background model can be structured using arithmetic mean of continuous images [39], median [40], or even by referring to histogram analysis result after a certain duration [41]. Simply put, basic modelling is easy to implement and is applicable to generate background models despite its shortcoming of failing to separate foreground by removing or introducing objects or when an object comes to a halt. Meanwhile, each pixel color distribution in statistical modelling turns into Gaussian distribution sum defined in the attributed color spaces. The widely

applied background subtraction algorithm refers to the parametric stochastic background model initiated by [42] and improved by [43]. Besides, single Gaussian, a combination of Gaussian, and Kernel Density Estimation [44] have been used in background modelling to distribute pixel color. Exceptional performance was displayed by Gaussian Mixture Models (GMMs) for outdoor scene analysis; thus, vastly applied with backgrounds in motion. The model also exhibits the capability to process changes despite poor lighting.

Unfortunately, the algorithm fails to give desired outcomes for radical changes as a consequence of motion in background, jittery camera, shadows, and changes in lighting. Besides, a background model becomes inefficient when established by noise-filled video frame at the learning phase. Many studies have attempted using GMMs to enhance background subtraction. To enhance system adaptability to changes of illumination [45], update modified equation [46], and investigate distribution of 3D multi-variable Gaussian, a technique was prescribed to calculate the number of ideal Gaussian distributions of every pixel automatically rather than fixing a constant [47]. Recently, a new framework was proposed by incorporating motion compensation and hysteresis thresholding [48]. Two background modelling; texture modelling and GMM, had successfully lowered cases of false positive. As for cluster background modelling, each frame pixel is presented based on time via clustering. The pixels are categorized based on if congruous cluster is a fraction of background, in comparison to relevant cluster group. In the clustering method, Codebook or K-mean algorithm is deployed. As for the fuzzy approach, the boundary of application varies substantially based on setting. Background subtraction was performed in a study based on similarity criterion with image and color of the input image [49]. Satisfactory outcomes obtained from Choquet Integral had several features: texture, color, and edge [50].

Meanwhile, background modelling based on neural network denotes the average of weights adequately trained neural networks for N number of clean frames. Training of the networks classifies every pixel into foreground or background [51, 52]. Culibrk et al., prescribed an approach of segmentation using the adaptive form of Probabilistic Neural Networks (PNNs) [53], while Maddalena and Petrosino deployed the Self-Organising Map (SOM) network in order to execute background subtraction [54]. The latter further enhanced their past study by incorporating the fuzzy function at the learning phase [55].

2.3 Object detection

The field of object detection has displayed substantial progress using deep learning in its applications. It differs from object recognition because the latter classifies every image into a pre-defined class and the former detects in every image via localization. Many object detection approaches using deep learning deployed CNNs [56–57]. From the (Regions with CNN) R-CNN advent [58], the combination of CNN classification and region proposal emerged as the preferred object detection framework. Rather than applying hand-crafted features (e.g., histograms of oriented gradients (HoG)) [59], CNN features are employed in R-CNN to offer better representation. As for R-CNN, thousands of bounding boxes or known as ‘region proposals’ have been generated using selective search, whereby the proposals turned into CNN classification inputs. Rapid R-CNN [60] is complemented by R-CNN for accuracy and efficiency. Initially, the

proposals share weights of forward pass in CNN via ROI pooling approach to minimize computation. Besides, bounding box regressor, features of convolutional, and classifier are linked in a network for a speedy system. Nonetheless, the selective search for region proposals is not efficient still.

Rapid R-CNN [61] was enhanced with the integration of region proposal process via detection networks. Region proposal network and convolutional layers are applied for proposals creation. Although the process of detection turns more rapid than Fast R-CNN, its speed lags from real-time (5 fps on GPU). The YOLO [62] – a cutting-edge detection technique – has surpassed the above techniques. Based on CNN, YOLO applies different framework. This new method divides an image into grid cells and estimates both each cell probabilities and bounding boxes coordinates. Probabilities are aggregated by calculating individual box confidence. The processing time can be increased substantially by this framework; images are processed at 40-90 fps on GPU and a tiny version exceeds 200 fps.

2.4 Pedestrian detection

One crucial task of video surveillance is pedestrian detection, which has two groups: learning-based and hand-crafted feature-based techniques. The latter category method of HoG detects images using shapes and local object appearance [63]. To date, HoG is the baseline used to extend algorithms. For instance, [64] combined HoG and LBP to address occlusion issue in pedestrian detection and reached 91.3% rate of detection using INRIA dataset [65]. Besides, a model based on part was deployed to enhance algorithms accuracy and detection efficiency [66]. Meanwhile, (HoG + LUV) color channels have been applied vastly [67-68]. Pedestrian detection also applies Haar-like characteristics. An efficient and simple informed-Haar detector was initiated by [69] for persons standing upright. Outcomes retrieved from Caltech and INRIA datasets [70] revealed 34.6% and 14.43% miss rates, respectively. Subjects in INRIA stood upright with high quality, whereas Caltech is a benchmark with challenging images [71]. Some techniques involved were Support Vector Machine (SVM) [72] and boosted classifiers [73]. Other frameworks applied hand-crafted features, e.g., HSG-HIK [74]. Recently, learning-based techniques have garnered much attention from the academic. In these approaches, image pixels are learnt, and the cutting-edge detection approaches rely on deep CNNs [75]. Another study used unsupervised convolutional sparse auto-encoders to train features, while classification deployed end-to-end supervised training [76]. The use of ConvNet with multi-stage features and INRIA dataset led to 10.55% average error rate. Meanwhile, unified CNN-based deep model was introduced using learning process to allow interactions among classification elements, feature extraction, occlusion, and part deformation [77].

In coping with intricate variations in pedestrian appearance, TA-CNN initiated by [78] optimized pedestrian classification using auxiliary semantic tasks (scene & pedestrian features), besides reducing miss rates in ETH and Caltech datasets [63]. Next, the approach of DeepParts [79] uses part detectors to solve occlusion issue, thus minimizing miss rate to 11.89% in Caltech dataset. Person detection also employed R-CNN, which can offer 53.9% human classification accuracy in VOC2011 dataset [80], whereas methods of other region (e.g., Regions and Parts) offer lower

accuracy. The MixedPeds algorithm [80] generates mixed reality dataset that amalgamates synthetic human agents and actual background. Although Faster R-CNN enhances the precision of detection, it demands substantial time and memory – minimizing the speed of detection**.**

3. CHALLENGES AND ISSUES:

The three conditions that ascertain excellent function of background subtraction are: fixed camera, constant illumination, and static background (pixels with unimodal distribution and no background object inserted or moved). Such optimum setting promotes background subtraction to offer excellent outcomes. In actual reality, some elements may disrupt the process. In 1999, Toyama et al. [81] outlined 10 challenging conditions for video surveillance. The list is extended to 13 in this paper:

- *Noisy image*: Images with poor quality, e.g., images from web cam or post-compression.
- Camera jitter: Swaying camera, thus leading to nominal motion in sequence. Foreground mask gives false detection as a result of motion with no proper maintenance mechanism.
- *Camera automatic adjustments*: Modern cameras have brightness and gain control, focus, and white balance that are automatic. Such adjustments alter the dynamic of color levels between frames in sequence.
- *Illumination changes*: Gradual (outdoor scene) or sudden (indoor light switch). Figure 1 illustrates indoor setting that denotes gradual change of illumination. This leads to false detection in some fractions of foreground mask (see Figure 1). In some cases, sudden change is portrayed in illumination as a result of light being turned on and off. Since all pixels are impacted by those changes, numerous false detections are identified.
- *Bootstrapping*: In training phase, background is unavailable for certain settings. This, it is not possible for computing background image representative.
- *Camouflage*: Features of foreground objects pixel can be subsumed by background modelling. Thus, background and foreground cannot be differentiated.
- *Foreground aperture*: Changes in moving object with uniform-colored regions cannot be detected and the whole object would not emerge as foreground. Foreground masks have false negative detections.
- *Moved background objects*: Background objects that move are not part of the foreground. No robust maintenance mechanism used to detect initial and new object positions.
- *Inserted background objects*: Insertion of a new background object, but not part of foreground. No robust maintenance mechanism used to detect the inserted background object.
- *Dynamic backgrounds*: Backgrounds can vacillate, and this demands models that represent disjoint sets of pixel values. There are three dynamic backgrounds: water surface, waving trees, and water rippling. Huge number of false detections is noted for every case.
- *Beginning moving object*: When the object moves first in the background, both it and new parts in the background (ghost) can be identified.

- *Sleeping foreground object*: Foreground object that stops moving cannot be differentiated from background object and incorporated in the background. Managing this scenario is context reliant. At times, foreground objects that stop moving must be included, but otherwise in some.
- *Shadows*: A foreground deriving from objects in motion or background.

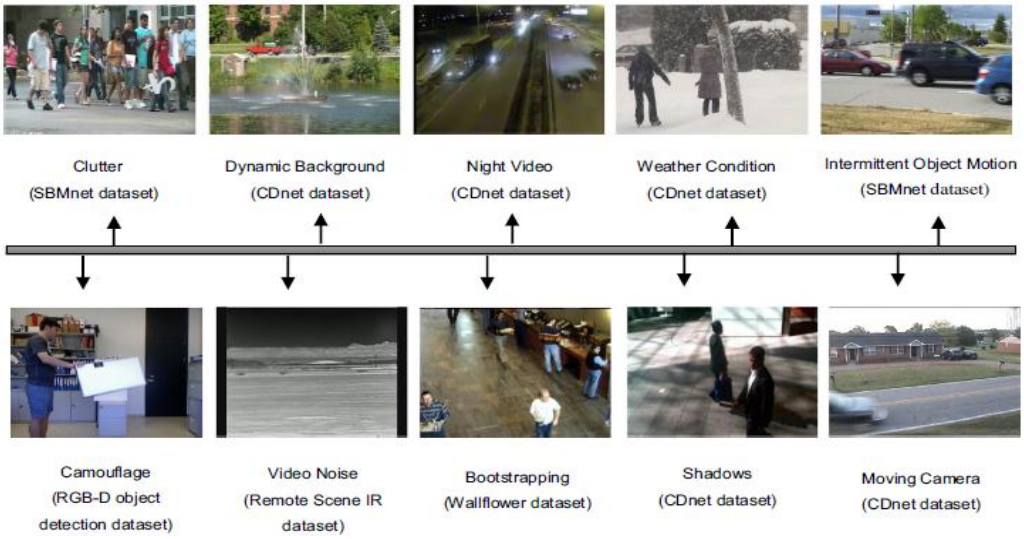

Figure 1: Challenges of moving objects detection

4. PROPOSED DETECTION MODEL OF MOVING OBJECTS IN VIDEOS

In this section, an enhanced method resulting from hybridizing the weight optimizations from CNN with a set of customized features derived by the Viola- Jones detector to enhance the overall network's performance, as depicted in Figure 2. In detail, the first convolutional layer of CNN is replaced with a function using customized filters, in order to minimize the required operations to calculate the layer's outputs. The replaced filters are composed of four rectangles within the filter space. These filters provide optimal computations of feature values than a conventional convolutional layer, when combined with an integral image. The initial experiments revealed that the customized filters of shape 5× 5 need only about 64% of the operations that a conventional convolutional layer requires. Empirically, these layers' real-world processing time behaves as theoretically computed and the optimization does not significantly reduce accuracy.

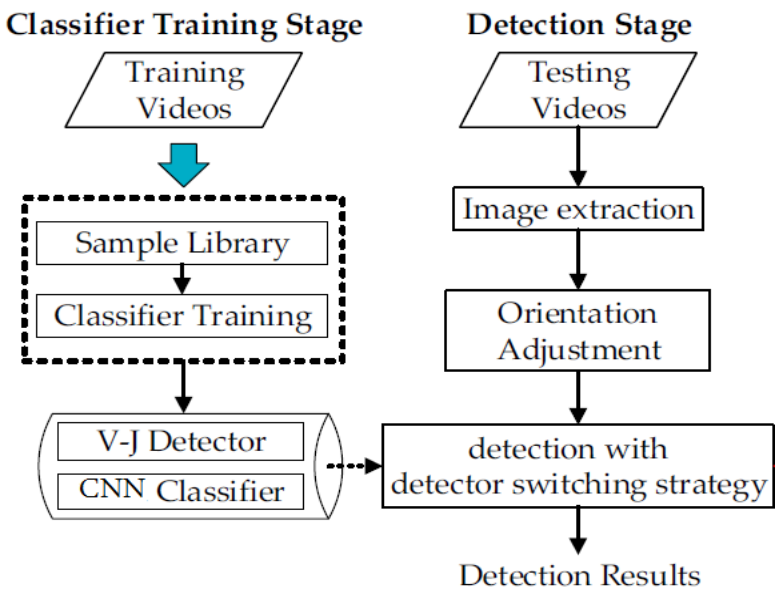

Figure 2: Proposed Detection Model

To determine the effectiveness of the proposed model, CDnet 2014 dataset is used to train and test the models. The initial results reveal that the proposed model outperform the state-of-the-art models (DMFC3D [84], Cascade [82], DeepBS [83]) in terms of accuracy, precision, recall and F-measure, as illustrated in Figure 3.

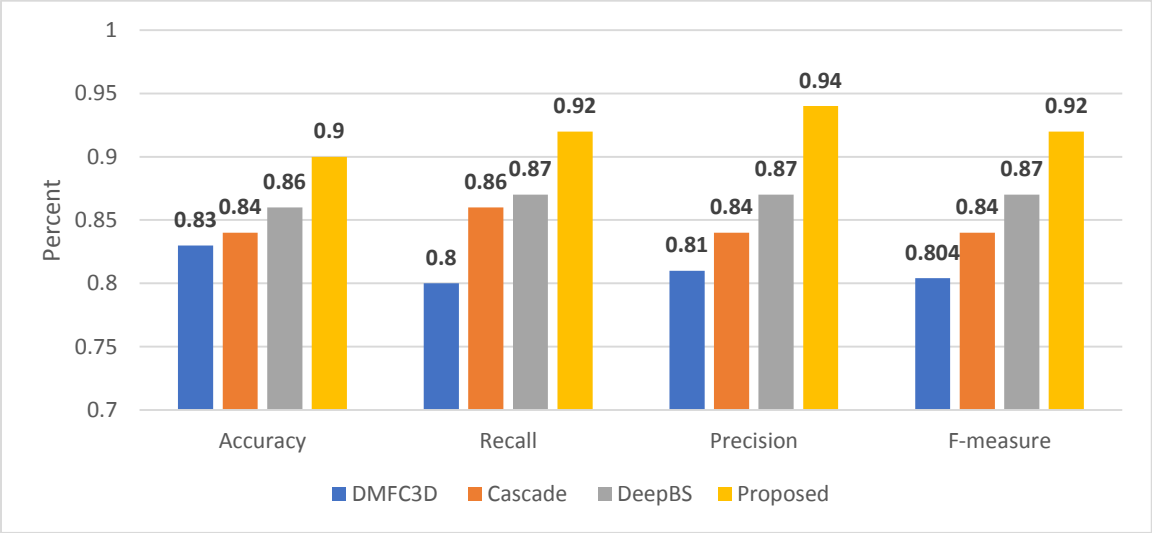

Figure 3: Evaluation results.

5. CONCLUSION:

This study has combed through recent and conventional background models applied for background subtraction. Two vital features were identified: (1) Background models classification based on mathematical tools, which is absent in the literature and (2) resources (recent & conventional libraries & datasets). Thus, future endeavor should look into sparse and RPCA models that display great potential to separate background from foreground in both actual and

incremental deployment. Fuzzy models too may be assessed for their potential and feature selection has emerged as an issue within this domain. In this article, an enhanced method is proposed, resulting from hybridizing the weight optimizations from CNN with a set of customized features derived by the Viola- Jones detector to enhance the overall network's performance. The initial findings show that the proposed method superior the state-of-the-art models. In the future, detailed stage of the proposed model will be further discussed in verified, and intensive experiments will be conducted to ensure its effectiveness and validity.

REFERENCES:

[1] S. Cheung, C. Kamath, Robust background subtraction with foreground validation for urban traffic video, EURASIP J. Appl. Signal Process. (2005).

[2] Y. Tian, A. Senior, M. Lu, Robust and efficient foreground analysis in complex surveillance videos, Mach. Vis. Appl. 23 (5) (2012) 967–983.

[3] A. Senior, Y. Tian, M. Lu, Interactive motion analysis for video surveillance and long-term scene monitoring, in: Asian Conference on Computer Vision, ACCV 2010 Workshops, 2010, pp. 164–174.

[4] F. El Baf, T. Bouwmans, Comparison of background subtraction methods for a multimedia learning space, in: International Conference on Signal Processing and Multimedia, SIGMAP, July 2007.

[5] J. Carranza, C. Theobalt, M. Magnor, H. Seidel, Freeviewpoint video of human actors, ACM Trans. Graph. 22 (3) (2003) 569–577.

[6] S. Elhabian, K. El-Sayed, S. Ahmed, Moving object detection in spatial domain using background removal techniques - state-of-art, Recent Patents Comput. Sci. 1 (1) (2008) 32–54.

[7] M. Cristani, M. Farenzena, D. Bloisi, V. Murino, Background subtraction for automated multisensor surveillance: A comprehensive review, EURASIP J. Adv. Signal Process. (2010) 24.

[8] T. Bouwmans, F. El-Baf, B. Vachon, Statistical background modeling for foreground detection: A survey, in: Handbook of Pattern Recognition and Computer Vision, vol. 4(2),World
Scientific Publishing, 2010, pp. 181–199.

[9] R. Radke, S. Andra, O. Al-Kofahi, B. Roysam, Image change detection algorithms: a systematic survey, IEEE Trans. Image Process. 14 (3) (2005) 294–307.

[10] del Postigo, C.G., Torres, J., Menéndez, J.M.: Vacant parking area estimation through background subtraction and transience map analysis. IET Intel. Transp. Syst. 9(9), 835–841 (2015)

[11] Muniruzzaman, S., Haque, N., Rahman, F., Siam, M., Musabbir, R., Hadiuzzaman, M., Hossain, S.: Deterministic algorithm for traffic detection in free-flow and congestion using video sensor. J. Built. Environ. Technol. Eng. 1, 111–130 (2016).

[12] Penciuc, D., El Baf, F., Bouwmans, T.: Comparison of background subtraction methods for an interactive learning space. NETTIES 2006 (2006)

[13] Zhang, X., Tian, Y., Huang, T., Dong, S., Gao, W.: Optimizing the hierarchical prediction and coding in HEVC for surveillance and conference videos with background modeling. IEEE Trans. Image Process. 23(10), 4511–4526 (2014).

[14] Bansod, S.D., Nandedkar, A.V.: Crowd anomaly detection and localization using histogram of magnitude and momentum. Vis. Comput. 36(3), 609–620 (2020)

[15] Mukherjee, S., Gil, S., Ray, N.: Unique people count from monocular videos. Vis. Comput. 31(10), 1405–1417 (2015).

[16] Huang, H., Fang, X., Ye, Y., Zhang, S., Rosin, P.L.: Practical automatic background substitution for live video. Comput. Vis. Media 3(3), 273–284 (2017)

[17] Tamás, B.: Detecting and analyzing rowing motion in videos. In: BME Scientific Student Conference (pp. 1–29) (2016)
[18] Zivkovic, Z., Van Der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn. Lett. 27(7), 773–780 (2006)
[19] Huang, W., Zeng, Q., Chen, M.: Motion characteristics estimation of animals in video surveillance. In: Proceedings of the 2017 IEEE 2nd Advanced Information Technology, Electronic
and Automation Control Conference (IAEAC) (pp. 1098–1102). IEEE (2017)
[20] Giraldo-Zuluaga, J. H., Salazar, A., Gomez, A., Diaz-Pulido, A.: Automatic recognition of mammal genera on camera-trap images using multi-layer robust principal component analysis and mixture neural networks (2017)
[21] Yang, Y., Yang, J., Liu, L., Wu, N.: High-speed target tracking system based on a hierarchical parallel vision processor and graylevel LBP algorithm. IEEE Trans Syst Man Cybern Syst 47(6),
950–964 (2016)
[22] Hadi, R.A., George, L.E., Mohammed, M.J.: A computationally economic novel approach for real-time moving multi-vehicle detection and tracking toward efficient traffic surveillance. Arab J Sci Eng 42(2), 817–831 (2017)
[23] Choudhury, S.K., Sa, P.K., Bakshi, S., Majhi, B.: An evaluation of background subtraction for object detection vis-a-vis mitigating challenging scenarios. IEEE Access 4, 6133–6150 (2016).
[24] Chapel, M.N., Bouwmans, T.: Moving objects detection with a moving camera: a comprehensive review. Comput. Sci. Rev. 38, 100310 (2020)
[25] Bouwmans, T.: Traditional and recent approaches in background modeling for foreground detection: an overview. Comput. Sci. Rev. 11, 31–66 (2014)
[26] Mabrouk AB, Zagrouba E. Abnormal behavior recognition for intelligent video surveillance systems: a review. Expert Syst Appl. 2018;91:480–91.
[27] Foroughi H, Aski BS, Pourreza H. Intelligent video surveillance for monitoring fall detection of elderly in home environments. In: 11th international conference on computer and information technology, 2008. ICCIT 2008. New York: IEEE; 2008. p. 219–24.
[28] Lao W, Han J, De With PH. Automatic video-based human motion analyzer for consumer surveillance system. IEEE Trans Consum Electron. 2009;55(2):591–8.
[28] Chen DY, Huang PC. Motion-based unusual event detection in human crowds. J Vis Commun Image Represent. 2011;22(2):178–86.
[29] Arroyo R, Yebes JJ, Bergasa LM, Daza IG, Almazán J. Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls. Expert Syst Appl. 2015;42(21):7991–8005.
[30] Sidhu RS, Sharad M. Smart surveillance system for detecting interpersonal crime. In: 2016 International Conference on communication and signal processing (ICCSP). New York: IEEE; 2016. p. 2003–7.
[31] Valera M, Velastin SA. Intelligent distributed surveillance systems: a review. IEEE Proc Vis Image Signal Process. 2005;152(2):192–204.
[32] Conde C, Moctezuma D, De Diego IM, Cabello E. Hogg: Gabor and hog-based human detection for surveillance in non-controlled environments. Neurocomputing. 2013;100:19–30.
[33] Huang K, Wang L, Tan T, Maybank S. A real-time object detecting and tracking system for outdoor night surveillance. Pattern Recog. 2008;41(1):432–44.
[34] Toyama K, Krumm J, Brumitt B, Meyers B. Wallflower: principles and practice of background maintenance. In: The Proceedings of the seventh IEEE international conference on computer vision, 1999, vol. 1. New York: IEEE; 1999. p. 255–61.

[35] Sobral A, Vacavant A. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. Comput Vis Image Underst. 2014;122:4–21.
[36] Bouwmans T. Background subtraction for visual surveillance: a fuzzy approach. Handb Soft Comput Video Surveill. 2012;5:103–38.
[37] Lee B, Hedley M. Background estimation for video surveillance. In: Image & Vision Computing New Zealand (IVCNZ '02). Auckland, NZ; 2002. p. 315–20.
[38] McFarlane NJ, Schofield CP. Segmentation and tracking of piglets in images. Mach Vis Appl. 1995;8(3):187–93.
[39] Zheng J, Wang Y, Nihan N, Hallenbeck M. Extracting roadway background image: mode-based approach. Transp Res Rec J Transp ResBoard. 1944;82–88:2006.
[40] Stauffer C, Grimson WEL. Adaptive background mixture models for real-time tracking. In: IEEE computer society conference on computer vision and pattern recognition, vol. 2. New York: IEEE; 1999. p. 246–52.
[41] Hayman E, Eklundh JO. Statistical background subtraction for a mobile observer. In: Proceedings of the international conference on computer vision. New York: IEEE; 2003. p. 67–74.
[42] Elgammal A, Harwood D, Davis L. Non-parametric model for background subtraction. In: Proceedings of the European conference on computer vision. Berlin: Springer; 2000. p. 751–67.
[43] Kaewtrakulpong P, Bowden R. An improved adaptive background mixture model for realtime tracking with shadow detection. In: Proceedings of 2nd European workshop on advanced video-based surveillance systems. Dordrecht: Brunel University; 2001.
[44] Conaire C, Cooke E, O'Connor N, Murphy N, Smearson A. Background modelling in infrared and visible spectrum video for people tracking. In: CVPR'05 Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition—workshops. CVPR workshops. New York: IEEE; 2005. p. 20.
[45] Zivkovic Z, Van Der Heijden F. Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn Lett. 2006;27(7):773–80.
[46] Yeh C-H, Lin C-Y, Muchtar K, Lai H-E, Sun M-T. Three-pronged compensation and hysteresis thresholding for moving object detection in real-time video surveillance. IEEE Trans Ind Electron. 2017;64:4945–55.
[47] Zhang H, Xu D. Fusing color and texture features for background model. In: Proceedings 3 of the third international conference fuzzy systems and knowledge discovery, FSKD 2006, Xi'an, China, September 24–28, 2006. Berlin: Springer; 2006. p. 887–93.
[48] El Baf F, Bouwmans T, Vachon B. Foreground detection using the choquet integral. In: WIAMIS'08 Proceedings of the 2008 ninth international workshop on image analysis for multimedia interactive services. New York: IEEE; 2008. p. 187–90.
[49] Culibrk D, Marques O, Socek D, Kalva H, Furht B. Neural network approach to background modeling for video object segmentation. IEEE Trans Neural Netw. 2007;18(6):1614–27.
[50] Bouwmans T. Recent advanced statistical background modeling for foreground detection—a systematic survey. Recent Pat Comput Sci. 2011;4(3):147–76.
[51] Maddalena L, Petrosino A. A self-organizing approach to background subtraction for visual surveillance applications. IEEE Trans Image Process. 2008;17(7):1168–77.
[52] Maddalena L, Petrosino A. A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection. Neural Comput Appl. 2010;19(2):179–86.
[53] Gkioxari G, Girshick RB, Malik J. Actions and attributes from wholes and parts; 2014. CoRR. abs/1412.2604.

[54] Kong T, Yao A, Chen Y, Sun F. Hypernet: towards accurate region proposal generation and joint object detection. In: The IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas, NV; 2016. p. 845–53.
[55] Yang F, Choi W, Lin Y. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In: The IEEE conference on computer vision and pattern recognition (CVPR); 2016.
[56] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014. p. 580–7.
[57] Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell. 2010;32(9):1627–45.
[58] Girshick R. Fast R-CNN. In: Proceedings of the IEEE international conference on computer vision. New York: IEEE; 2015. p. 1440–8.
[59] Ren S, He K, Girshick R, Sun J. Faster r-cnn: towards real-time object detection with region proposal networks. In: The conference on advances in neural information processing systems. Montréal: Curran Associates; 2015. p. 91–9.
[60] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017; 2017. p. 6517–25.
[61] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: CVPR'05 Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition. CVPR 2005, vol. 1. New York: IEEE; 2005. p. 886–93.
[62] Wang X, Han TX, Yan S. An hog-lbp human detector with partial occlusion handling. In: 2009 IEEE 12th international conference on computer vision. New York: IEEE; 2009. p. 32–9.
[63] Dollár P, Appel R, Belongie S, Perona P. Fast feature pyramids for object detection. IEEE Trans Pattern Anal Mach Intell. 2014;36(8):1532–45.
[64] Dollár P, Appel R, Kienzle W. Crosstalk cascades for frame-rate pedestrian detection. In: Proceedings of the 12th European conference on computer vision (ECCV) 2012. Berlin: Springer; 2012. p. 645–59.
[65] Zhang S, Bauckhage C, Cremers AB. Informed haar-like features improve pedestrian detection. In: 2014 IEEE conference on computer vision and pattern recognition. p. 947–54; 2014.
[66] Luo P, Tian Y, Wang X, Tang X. Switchable deep network for pedestrian detection. In: 2014 IEEE conference on computer vision and pattern recognition; 2014. p. 899–906.
[67] Benenson R, Omran M, Hosang JH, Schiele B. Ten years of pedestrian detection, what have we learned? 2014. CoRR, abs/1411.4304.
[68] Maji S, Berg AC, Malik J. Classification using intersection kernel support vector machines is efficient. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. New York: IEEE; 2008. p. 1–8.
[69] Dollár P, Tu Z, Perona P, Belongie S. Integral channel features. In: Cavallaro A, Prince S, Alexander D, editors. Proceedings of the British Machine Vision Conference. BMVA Press; 2009. p. 91.1–11.
[70] Bilal M, Khan A, Khan MUK, Kyung CM. A low-complexity pedestrian detection framework for smart video surveillance systems. IEEE Trans Circuits Syst Video Technol. 2016;27:2260–73
[71] Kang K, Ouyang W, Li H, Wang X. Object detection from video tubelets with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 817–25.

[72] Sermanet P, Kavukcuoglu K, Chintala S, LeCun Y. Pedestrian detection with unsupervised multi-stage feature learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2013. p. 3626–33.
[73] Ouyang W, Wang X. Joint deep learning for pedestrian detection. In: Proceedings of the IEEE international conference on computer vision; 2013. p. 2056–63.
[74] Tian Y, Luo P, Wang X, Tang X. Pedestrian detection aided by deep learning semantic tasks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 5079–87.
[75] Luo P, Wang X, Tang X. Pedestrian parsing via deep decompositional network. In: 2013 IEEE international conference on computer vision; 2013. p. 2648–55.
[76] Tian Y, Luo P, Wang X, Tang X. Deep learning strong parts for pedestrian detection. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 1904–12.
[78] Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The PASCAL visual object classes challenge 2011 (VOC2011) results. http://host.robot s.ox.ac.uk/pasca l/VOC/voc20 11/resul ts/index .html.
[79] Arbeláez P, Hariharan B, Gu C, Gupta S, Bourdev L, Malik J. Semantic segmentation using regions and parts. In: 2012 IEEE conference on computer vision and pattern recognition; 2012. p. 3378–85.
[80] Cheung E, Wong A, Bera A, Manocha D. Mixedpeds: pedestrian detection in unannotated videos using synthetically generated human-agents for training. In: Proceedings of the AAAI conference on artificial intelligence. New Orleans, Louisiana, USA; 2018.
[81] Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: principles and practice of background maintenance. In: Proceedings of the Seventh IEEE International Conference on Computer Vision (Vol. 1, pp. 255–261). IEEE,1999.
[82] Y. Wang, Z., Luo & P., Jodoin,: Interactive deep learning method for segmenting moving objects. Pattern Recognition Letters, 96, 66-75, 2017.
[83] Babaee, M., Dinh, D. T., & Rigoll, G.: A deep convolutional neural network for video sequence background subtraction. Pattern Recognition, 76, 635-649, 2018.
[84] Wang, Y., Yu, Z., & Zhu, L: Foreground detection with deeply learned multi-scale spatial-temporal features. Sensors, 18(12), 4269, 2018.