An end-to-end deep learning approach for simultaneous background modeling and subtraction

V. Mondéjar-Guerra[12]
v.mondejar@udc.es
J. Rouco[12]
jrouco@udc.es
J. Novo[12]
jnovo@udc.es
M. Ortega[12]
mortega@udc.es

[1] Department of Computer Science, University of A Coruña, A Coruña, Spain

[2] CITIC-Research Center of Information and Communication Technologies, University of A Coruña, A Coruña, Spain

Abstract

Background subtraction is an active research topic due to its great utility on many video analysis applications. In this work, a new approach for background subtraction employing an end-to-end deep learning architecture is proposed. The proposed architecture consists in two nested networks that are trained together. The first one extracts the background model features of the scene from a small group of frames. The second performs the subtraction operation given the previous features and a target frame. In contrast to most of the recent deep learning proposals, our trained model can be used on any scene without the need of being retrained. The method has been trained and evaluated using the public CDnet2014 database following a scene-wise cross-validation approach. The obtained results show a competitive performance of the proposed method on background subtraction, proving its ability to extrapolate to unseen scenes.

1 Introduction

The separation of the moving objects (foreground) from the static context (background) on video sequences constitutes an important issue that is frequently required for many computer vision applications, like traffic video surveillance [11] or human tracking systems [17]. This task, known as background subtraction or foreground detection, has been an active research topic for more than 20 years [7]. The existent methods are usually divided in three main steps: background initialization, foreground detection, and background maintenance [3]. In the first step, a background model is generated from a set of representative images. Ideally, these images are free of foreground and include enough variability of any particularity of the scene, *e.g.* small movements of tree leaves due to wind or illumination changes. However, these ideal conditions can not always be satisfied, motivating the recent proposal of a large variety of approaches for background generation that can deal with the worst circumstances. These approaches range from basic operations like the median [13], to more

complex techniques like Mixtures of Gaussians (MOG) [5], Robust Principal Component Analysis (RPCA) [6], clustering [9], neural networks [10], and most recently, deep learning [7]. Next, during the foreground detection step, the similarity between a target frame and the reference background model is assessed to classify each pixel as background or foreground. Finally, many methods also include a maintenance step with the aim of updating the background model with changes, such as the addition of new static objects in the scene [27].

Most of the learning-based background subtraction proposals are trained in a supervised manner [7], *i.e.*, the desired segmented foreground maps (ground truth) are required during the training phase in order to adjust the system through the minimization of a loss function. In addition, it can be distinguished between scene-specific and non scene-specific methods. The former need to be retrained adjusting its weights for each particular scene, while the latter provide a generic system that can be directly used in multiple scenes. In general, scene-specific methods achieve better results because they receive more precise information during the training. However, the fact that a method is both scene-specific and supervised imposes an important limitation for many real world applications, since it would require the manual labeling of accurate foreground maps for each target scene, which is a tedious task.

In this work, a non scene-specific deep learning architecture, inspired by the classical background subtraction algorithms, is proposed. The proposed architecture simultaneously performs the background modeling and subtraction steps using two nested networks in an end-to-end process. The first network receives a small group of frames and computes a background model for the scene. Next, a target frame along with the previous computed background model feed the second network, which returns the background-foreground segmentation map. The full architecture is trained in a supervised way, by providing the ground truth for the target frames only and without explicitly providing the desired background model. This end-to-end training using frames of different scenes allows the learning of an unsupervised background modelling strategy (first network) and a generic foreground detection strategy (second network) using the generated scene model. This should allow the learned models to extrapolate to scenes that have not been seen by the networks during the training. At the best of our knowledge, no end-to-end deep learning solution simultaneously performing both the background modeling and subtraction tasks has been proposed before. Additionally, the existent deep learning approaches either require supervised scene-specific adjustment, using segmentation ground truth, or do not follow an experimentation strategy allowing to evaluate if the adjusted models extrapolate to unseen scenes.

2 Related Works

During the last decades, background subtraction solutions have put special attention on statistical distribution approximation approaches [11][25][5] and decomposition techniques [19][6]. Friedman and Russel were the first to propose unsupervised background modelling using a MOG [11]. Later, Stauffer and Gimson [25] added a background modelling updating strategy to improve its results in online settings. Since then, many other authors have proposed improvements to the mixture distribution approaches [5]. Alternatively, since the work of Oliver *et al.* [19], the unsupervised decomposition into low-rank plus sparse matrices has also been proved to be an appropriate solution to separate the foreground [6]. The success of these methods demonstrate that is possible to generate robust background models in an unsupervised way.

On the other hand, deep learning methods have outperformed many computer vision

problems during the last decade [15]. However, the improvement is often limited to applications where a supervised training is feasible, which requires a huge amount of properly annotated data. Recently, the existence of large datasets like CDnet2012 [12] and its extension CDnet2014 [29], which provide accurate ground truth for more than 100,000 frames, enabled a huge variety of works applying deep learning techniques on this area [7].

Some of these approaches are scene-specific solutions [8][16], requiring to be trained for each target scene. Conversely, other proposals are designed to, in principle, handle unseen scenes once they are trained [1][13][21]. Independently of this, the authors of some of these methods consider that, in order to perform background subtraction, a complex background modeling strategy is not required. In this sense, some works relegate the whole process to a deep learning architecture that analyses a temporal sequence of frames [13][21] or just a single target frame [16]. Braham and Van Droogenbroeck [8] proposed a network that receives a target frame along with a background image, which is computed as the temporal median of a group of frames, while Babaee *et al.* [1] compute an additional background model using a combination of the SuBSENSE [23] and Flux Tensor [28] algorithms. Our deep learning approach, instead, includes an explicit background modeling network that is trained along with the background subtraction in an end-to-end architecture.

Currently, scene-specific deep learning methods are dominating the performance of ranking over CDnet datasets[1], but they use part of the ground truth of the evaluated scenes during the training. In addition, neither of the previous deep learning approaches evaluate their performance under a mutually exclusive scene partition on the training and testing sets. Therefore, their extrapolation capabilities are not properly evaluated.

3 Proposed method

We designed an end-to-end architecture that performs in a single forward step both the background modeling and background subtraction tasks. A general overview of the proposed method is depicted in Fig. 1. As it can be observed, our architecture consists of two nested networks. The first network, denoted as Background Modeler Network (BMN), receives

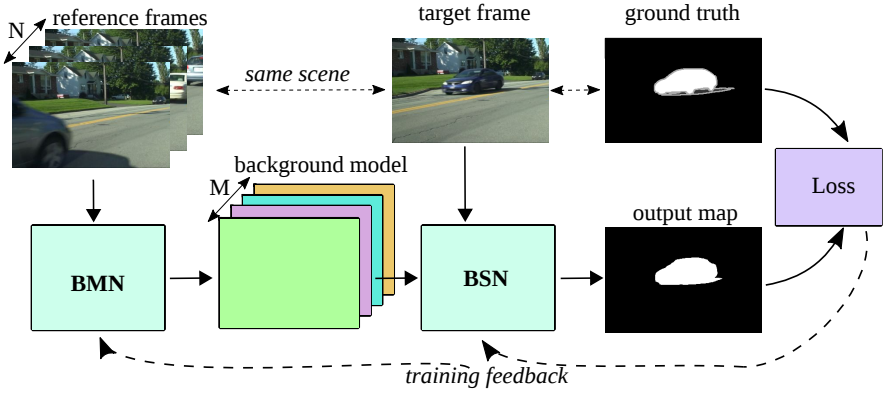

Figure 1: Overview diagram of the proposed BMN-BSN paradigm composed by the network cascade: background modeling network (BMN) and background subtraction network (BSN).

[1](see Results section in http://changedetection.net/)

N reference frames from a scene and generates a multidimensional map that represents the background model for that scene. These reference frames can present certain level of foreground, and summarize all the information that the network is allowed to see from a particular scene. Next, the obtained background model along with a target frame constitute the input of the second network, the Background Subtractor Network (BSN), which returns the segmented output map for that target frame. The network cascade is trained end-to-end with a supervised segmentation loss over the output map and the ground truth for the target frame. The desired background model for the scene is not explicitly enforced. Instead, the background modeling is learnt as a byproduct of the end-to-end training with multiple scenes, and with the only restriction that the input frames of both networks (reference and target frames) belong to the same scene.

3.1 Network architecture

Any network architecture that allows predicting full resolution multidimensional maps from input images may be suitable for being used as the BMN and the BSN. In this work, in particular, we employed an U-Net [20] architecture for both networks. This choice is motivated by the flexibility and simplicity of this architecture, its ability to integrate multiscale patterns into full resolution output maps, and its recent success in several segmentation tasks. Take in mind that this work aims at providing preliminary results using the proposed end-to-end paradigm for background modeling and subtraction. Thus, the selection of the optimal network architecture for this purpose is out of the scope of this work.

The U-Net (see Fig. 2) features a symmetric fully-convolutional encoder-decoder architecture with concatenated skip connections. The encoder is composed of several downsampling blocks that use VGG-like stacked 3×3 convolutional layers, with rectified linear unit (ReLU) activations followed by a 2×2 Max Pooling [22]. The number of convolutional feature channels is doubled on each subsequent block, starting from K. The decoder, instead, is composed of a sequence of upsampling blocks that are also connected to the corresponding encoder block (of the same resolution) through skip connections. To that end, the previous feature maps are upsampled using up-convolution operations and concatenated with the skipped feature maps. The rest of the upsampling block consist of a stack of ReLU convolutions, similar to that of the encoder's blocks. In the case of the decoder, the feature channels are decreased by a half on each block. The network output is composed of T 1×1 convolutions, depending on the desired number of output channels, while the number of input channels is controlled by the size of input maps.

The BMN takes N frames as input, resulting in $N \times 3$ channels assuming *RGB* inputs, and M background model channels as output. The BSN, instead, takes the M background model channels plus the target frame channels (3, in case of *RGB*) as input, and 2 output channels tied with a softmax activation, corresponding to the background and foreground classes. The used loss is normalized cross-entropy.

4 Experimentation and results

The public CDnet2014 database [29] was used to train and evaluate the proposed method. This database is organized in ten categories: *Baseline (BSL), Dynamic Background (DMB), Camera Jitter (CJT), Shadow (SHD), Intermittent Object Motion (IOM), Thermal (THM), Challenging Weather (BDW), Low Frame-Rate (LFM), Night (NGT), Pan Tilt Zoom (PTZ)*,

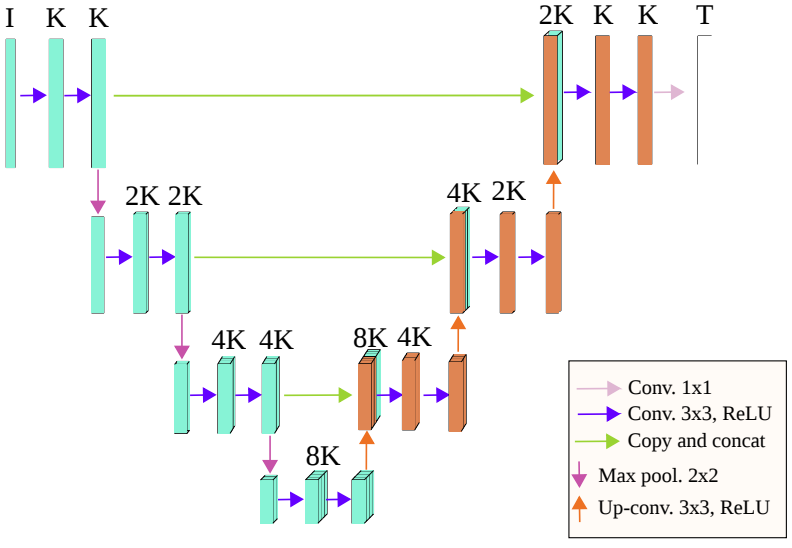

Figure 2: Diagram of the U-Net based architecture used for the BMN and the BSN.

and *Air Turbulence (TBL)*. Each category contains four or more videos captured on different scenes. The length of the videos ranges from 900 to 7,000 frames. The ground truth consist in 5 different values, representing: static (background), shadow, non-ROI, unknown, and moving (foreground) pixels. The unknown, non-ROI and shadow regions are not included for the loss computation during our experimentation. In this work, an offline background initialization approach is followed, thus the scenes from then PTZ category, which vary the camera viewpoint during the videos, are not considered. Each video of the database is divided in two parts. The ground truth is not available for the first part of the video (unlabelled set), while it is provided for the frames in the second part (labelled set). During our experimentation the frames from the unlabelled set are employed as reference frames to compute the background model, and those from the labelled set to perform the background subtraction.

4.1 Training Details

A scene-wise 3-fold cross-validation is applied in order to evaluate the capability of the proposed architecture to extrapolate to unseen scenes. The database is divided in three mutually exclusive video sets, trying to balance the number of frames between the different categories along the sets. Each set comprises about 30.000 frames. At each training iteration a target frame (input for the BSN) is selected from the labelled part of a video along with N randomly selected reference frames from the unlabelled part of the same video (inputs for the BMN). The target frames from different scenes are presented in random order to not overfit the network to any particular scene. A complete training epoch includes a forward pass of all the target frames from the training set. In our experimentation, each network was trained during 5 epochs. The Adam optimizer [14], with a learning rate of $1e^{-5}$, was used. The configuration of the network parameters N, M, and K was fixed to 16, 12, and 64, respectively. The input resolution for all the input images (scene frames, target frame and output

map) was fixed to 320×240 because is the most frequent resolution of the videos from the CDnet2014 database, but any other different resolution, or even a varying one, could have been selected, due to the fully-convolutional nature of the proposed U-Net cascade.

4.2 Results and Discussion

Table 1 shows the average results that were obtained by our method on videos from the different categories of CDnet2014. The evaluation metrics correspond with those included in CDnet2014 challenge [29]: Recall (Re), Specificity (Sp), False Positive Rate (FPR), False Negative Rate (FNR), Percentage of Wrong Classifications (PWC), Precision (Pr), and F-measure (F1 score). In general, our method achieved higher values for recall than for precision. That is, our method is not conservative, being more sensitive to detect most of the anomalies in the scene at the cost of introducing some FPs. The highest FPR is produced on the *IOM* category. That category is intended for testing how the algorithms adapt to background changes [29]. Given that our experimental setting follows an offline background initialization, assuming that the reference frames selected from the unlabelled part summarized the scene, and without performing any mechanism to update the model, a high FPR is expected in that category. Fig. 3 shows an example of FPs obtained by our method on a scene from the *IOM* category. In that scene, the car is static, being part of the background, during all the unlabelled set frames, but it changes its position during the labelled set frames. That fact causes the apparition of a huge number of FPs on the region in which the car originally was. The next highest FPR value is produced on the *CJT* category. This is also reasonable, considering the extra difficulties to perform background subtraction on videos that present unstable camera vibrations. The lowest F-measure and precision are achieved on the *TBL* category. In general, the anomalies (TP + TN) in videos from that category are small. That fact may explain the results on this category, considering that the apparition of FPs has more impact to decrease the precision (TP /(TP+FP)). For the *NGT* category, numerous FNs are caused by the lack of illumination, and many FPs are caused by headlight reflections [29]. The highest FNR is obtained for the *LFM* category. However, our method does not have *a priori* any inconvenience to process scenes from this category, as there is not a maintenance step, the reference frames are not required to be presented in order, and the proposed method does not rely on time related features. The fact that more FNs are produced on that cate-

(a) (b) (c)

Figure 3: Example of FPs produced by the proposed method on the *winterDriveway* video due to changes on the static background elements from the reference frames to the target frame. (a) Reference frame. (b) Target frame. (c) Segmented output of the BMN-BSN method. TP(Green), FP (Red), FN (Pink), TN (Black), Shadow, Non-ROI and Unknown (Gray levels).

Category	Re(%)	Sp(%)	FPR(%)	FNR(%)	PWC	F1(%)	Pr(%)
BDW	78.02	99.77	0.23	21.98	0.5355	81.24	85.31
BSL	96.77	99.73	0.27	3.23	0.3253	95.21	93.71
CJT	90.99	95.17	4.83	9.01	4.9312	69.62	60.69
DMB	86.76	98.21	1.79	13.24	1.8297	63.71	56.62
IOM	90.18	86.71	13.29	9.82	11.8070	63.69	57.14
LFM	65.38	99.82	0.18	34.62	0.9209	64.26	70.45
NGT	67.05	98.69	1.31	32.95	2.5853	61.25	61.13
SHD	96.70	98.87	1.13	3.30	1.2111	85.88	78.93
THM	75.34	98.99	1.01	24.66	1.6661	78.49	84.21
TBL	77.76	96.82	3.18	22.24	3.2476	55.44	55.03
Overall	82.50	97.28	2.72	17.51	2.9060	71.88	70.32

Table 1: Average results of the proposed method (BMN-BSN) over the videos of the different categories of CDnet2014 public database.

gory is related with other properties of those videos, like the illumination changes on the *port_0_17fps* video or the low contrast on the *turnpike_0_5fps* video.

Complementarily, Table 2 shows a comparison of the average F-measure (F1 score) obtained on each category by the proposed method and others from the literature. In this comparison we included some of the top ranked methods and some of the most commonly employed methods for background subtraction that were tested on the CDnet2014 database. FgSegnet_S [16], which is a supervised scene-specific deep learning method, obtains the best results by a large margin. Next, DeepBS [1] is a supervised deep learning method that relies on a combination of SuBSENSE [23] and Flux Tensor [28] to perform the background modeling, and applies a post-processing step on their outputs. It must be noted that these two methods use part of the ground truth from the evaluation scenes during the training. SuBSENSE [23] relies on color and local binary similarity patterns making pixel-level decisions with automatic online adjustment of tuning parameters. PAWCS [24] is a self adjusted online method that uses persistence-based words to model the background over longer periods of time. GMM-(S-G)[26] represents a consolidated online GMM updating approach, which is frequently cited and usually employed for real applications [4]. GMM-(S-G) obtained the lowest F-measure values for all the categories.

In general, our method presents a competitive performance in several categories of the database. Besides, despite the limited performance of the proposed method in categories like *IOM*, which are mostly motivated by the offline nature of our experimental settings, the obtained results are similar to other online methods in the literature. In addition, contrary to the other deep learning proposals, our method is tested employing a 3-fold cross validation scheme, which allows evaluating its extrapolation capabilities to unseen scenes without further training. Our method is capable of real-time processing running on a *GeForce RTX 2080 Ti*, while computing the background model maps (output of the BMN) at each forward step. As the background model maps could be computed just one time per scene, or at a limited rate of model updates, the throughput could be easily increased.

Fig. 4 shows an illustrative example of the background model generated by our network from 16 random frames of the video *PETS2006*. The background model, consisting in 12 maps, is represented as 4 RGB images. It can be observed that in spite of the scene frames not being free of foreground, the generated maps by the network only represents scene background features. Additionally, qualitative examples of the proposed method over scenes from

	Supervised					
	FgSegNet	DeepBS	SuBSENSE	PAWCS	**BMN-BSN**	GMM(S-G)
Category	[16]	[1]	[23]	[24]		[26]
BDW	**99.07**	83.01	**86.19**	81.52	81.24	73.80
BSL	**99.77**	95.80	95.03	93.97	**95.21**	82.45
CJT	**99.57**	89.90	**81.52**	81.37	69.62	59.69
DMB	**99.58**	87.61	81.77	**89.38**	63.71	63.30
IOM	**99.40**	60.98	65.69	**77.64**	63.69	52.07
LFM	**89.72**	60.02	64.45	**65.88**	64.26	53.73
NGT	**97.13**	58.35	55.99	41.52	**61.25**	40.97
SHD	**99.27**	93.04	**89.86**	89.13	85.88	73.70
THM	**99.37**	75.83	81.71	**83.24**	78.49	66.21
TBL	**96.81**	84.55	**77.92**	64.50	55.44	46.63
Overall	**97.97**	78.91	**78.01**	76.81	71.88	61.25

Table 2: Comparison of F-measure values, over the different categories from CDnet2014 database, achieved by the proposed method (BMN-BSN) and others from the literature. Supervised methods employed part of the ground truth of the evaluated scenes during training. In bold, best result per category.

the different categories of CDnet2014 database are showed on Fig. 5.

(a) (b)

Figure 4: Generated background model by the BMN. (a) $N = 16$ reference frames of the scene PETS2006. (b) Representation of the background maps ($M = 12$) as 4 images in RGB.

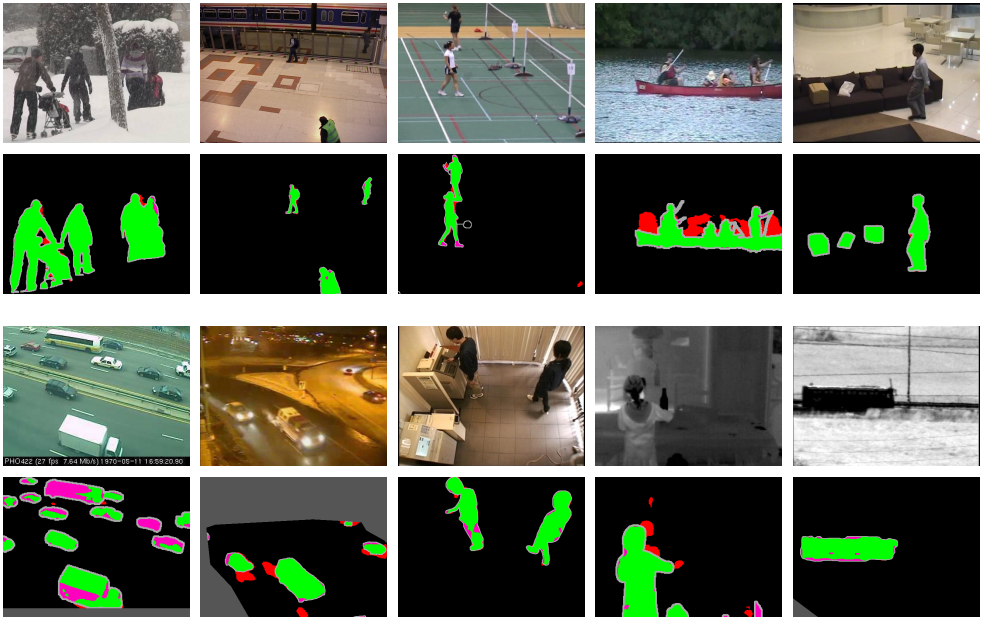

Figure 5: Qualitative results for background subtraction task obtained by our network from scenes of the different categories of CDnet2014 database. TP(Green), FP (Red), FN (Pink), TN (Black), Shadow, Non-ROI and Unknown (Gray levels).

5 Conclusions

A novel end-to-end deep learning approach that simultaneously performs background modeling and subtraction tasks is presented in this work. The proposed paradigm allows for the trained model to be used on unseen scenes without needing to be trained with scene-specific ground truth data. The proposed architecture is able to extract background model features from a small group of frames that may contain foreground, or be affected by adverse conditions like snowing or water motion. The network has been evaluated on the public CDnet2014 database employing a scene-wise cross-validation, evaluating the extrapolation capabilities of the network. The obtained results demonstrate the capabilities of the network to handle unseen scenes under a wide variety of situations. Despite the offline background initialization setting used during the experimentation, which contradicts the purpose of some categories of the CDnet2014 database (like the *IOM* category) the method demonstrates a competitive performance against other online methods in the literature.

Jittering videos still present a challenge on the proposed architecture. An additional step that aligns the background model features making a jitter invariant background modelling strategy should be studied as possible future works. On the other hand, we consider that background modeling is more complex than the subtraction operation, and having a good enough reference background model would allow a trivial subtraction operation. Therefore, different strategies to increase the robustness of the background modeling task should be studied and tested.

Acknowledgement

This work has received financial support from the Xunta de Galicia (Centro singular de investigación de Galicia accreditation 2016-2019) and the European Union (European Regional Development Fund - ERDF), PC18/01. Also, this work has received financial support from the ERDF and the Xunta de Galicia, Centro singular de investigación de Galicia accreditation 2016-2019, Ref.ED431G/01; and Grupos de Referencia Competitiva, Ref. ED431C 2016-047.

References

[1] M. Babaee, D. T. Dinh, and G. Rigoll. A deep convolutional neural network for video sequence background subtraction. *Pattern Recognition*, 76:635–649, 2018. ISSN 0031-3203.

[2] T. Bouwmans. Background Subtraction For Visual Surveillance: A Fuzzy Approach. In *Handbook on Soft Computing for Video Surveillance*, pages 103–134. CRC Press, 2012.

[3] T. Bouwmans. Traditional and recent approaches in background modeling for foreground detection: An overview. *Computer Science Review*, 11-12:31–66, 2014. ISSN 15740137.

[4] T. Bouwmans and B. Garcia-Garcia. Background Subtraction in Real Applications: Challenges, Current Models and Future Directions. 2019.

[5] T. Bouwmans, F. El Baf, and B. Vachon. Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey. *Recent Patents on Computer Sciencee*, 1(3):219–237, 2008. ISSN 22132759.

[6] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E. H. Zahzah. Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset. *Computer Science Review*, 23:1–71, 2017. ISSN 1574-0137.

[7] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung. Deep Neural Network Concepts for Background Subtraction: A Systematic Review and Comparative Evaluation. *CoRR*, abs/1811.05255, 2018.

[8] M. Braham and M. Van Droogenbroeck. Deep background subtraction with scene-specific convolutional neural networks. *International Conference on Systems, Signals, and Image Processing*, 2016-June(CDnet):3–6, 2016. ISSN 21578702.

[9] D. E. Butler, V. M. Bove, and S. Sridharan. Real-Time Adaptive Foreground/Background Segmentation. *EURASIP Journal on Advances in Signal Processing*, 2005(14):841926, 2005. ISSN 1687-6180.

[10] D. Culibrk, O. Marques, D. Socek, H. Kalva, and B. Furht. Neural Network Approach to Background Modeling for Video Object Segmentation. *IEEE Transactions on Neural Networks*, 18(6):1614–1627, 2007. ISSN 1045-9227.

[11] N. Friedman and S. Russell. Image Segmentation in Video Sequences: A Probabilistic Approach. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, UAI'97, pages 175–181. Morgan Kaufmann Publishers Inc., 1997. ISBN 1-55860-485-5.

[12] N. Goyette, P. Jodoin, F. Porikli, J. Konrad, and P. Ishwar. Changedetection.net: A new change detection benchmark dataset. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2012.

[13] Z. Hu, T. Turki, N. Phan, and J. T. L. Wang. A 3D atrous convolutional long short-term memory network for background subtraction. *IEEE Access*, 6:43450–43459, 2018. ISSN 21693536.

[14] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 5 2015. ISSN 0028-0836.

[16] L. A. Lim and H. Y. Keles. Foreground segmentation using convolutional neural networks for multiscale feature encoding. *Pattern Recognition Letters*, 112:256–262, 2018. ISSN 0167-8655.

[17] D. Makris and T. Ellis. Learning semantic scene models from observing activity in visual surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(3):397–408, 2005. ISSN 1083-4419.

[18] N. J. B. McFarlane and C. P. Schofield. Segmentation and tracking of piglets in images. *Machine Vision and Applications*, 8(3):187–193, 1995. ISSN 1432-1769.

[19] N. M. Oliver, B. Rosario, and A. P. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000. ISSN 0162-8828.

[20] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015. ISBN 978-3-319-24574-4.

[21] D. Sakkos, H. Liu, J. Han, and L. Shao. End-to-end video background subtraction with 3d convolutional neural networks. *Multimedia Tools and Applications*, 77(17): 23023–23041, 2018. ISSN 1573-7721.

[22] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[23] P. St-Charles, G. Bilodeau, and R. Bergevin. SuBSENSE: A Universal Change Detection Method With Local Adaptive Sensitivity. *IEEE Transactions on Image Processing*, 24(1):359–373, 2015. ISSN 1057-7149.

[24] P. St-Charles, G. Bilodeau, and R. Bergevin. A self-adjusting approach to change detection based on background word consensus. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 990–997, Jan 2015.

[25] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 2, pages 246–252 Vol. 2, 1999.

[26] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000. ISSN 0162-8828.

[27] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: principles and practice of background maintenance. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 255–261 vol.1, 1999.

[28] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan. Static and Moving Object Detection Using Flux Tensor with Split Gaussian Models. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 420–424, 2014.

[29] Y. Wang, P. M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar. CDnet 2014: An Expanded Change Detection Benchmark Dataset. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW '14, pages 393–400. IEEE Computer Society, 2014. ISBN 978-1-4799-4308-1.