Refinement Network for unsupervised on the scene Foreground Segmentation

Montse Pardàs
Signal Theory and Communications Dept
Universitat Piolitècnica de Catalunya
Barcelona, Spain
montse.pardas@upc.edu

Gemma Canet
Signal Theory and Communications Dept
Universitat Politècnica de Catalunya
Barcelona, Spain
gemmacanettarres@gmail.com

***Abstract*—In this paper we present a network for foreground segmentation based on background subtraction which does not require specific scene training. The network is built as a refinement step on top of classic state of the art background subtraction systems. In this way, the system combines the possibility to define application oriented specifications as background subtraction systems do, and the highly accurate object segmentation abilities of deep learning systems. The refinement system is based on a semantic segmentation network. The network is trained on a common database and is not fine-tuned for the specific scenes, unlike existing solutions for foreground segmentation based on CNNs. Experiments on available databases show top results among unsupervised methods.**

***Index Terms*—Background subtraction, semantic segmentation networks, refinement network**

I. Introduction

The process of identifying each pixel in an image as part of foreground or background is usually known as 'foreground segmentation'. This is widely employed for many applications such as detecting intruding objects in surveillance systems, automatically monitoring large and crowded areas such as airports, or traffic monitoring at roads. It is also a very useful tool for analysing human activities, and even editing videos for the cinematographic industry by changing or isolating their background.

This problem has been widely studied during the last decade. Classical methods are based on modelling the background per-pixel distribution and identifying the foreground pixels as those that are an exception to their corresponding background model. Recently, deep neural networks have been used, both for modeling the background and for the subtraction step. While outstanding results are obtained using neural networks, these methods use a training or fine-tuning step based on the sequence to evaluate. In this paper, we explore the possibility of using neural networks for background subtraction without need for training or fine-tuning on the testing scenes, as in real scenarios it is not feasible to require a user-provided detailed segmentation for a usually high number of frames, or even to devote a few hours to train the network with non labeled background.

For this, we propose to use a two-steps scheme. The first step uses a conventional foreground detection that produces a rough approximation to the foreground. These conventional systems have the possibility to accommodate to different requirements of the application scenario, such as strong shadows, multi-modal background, or delay in incorporating moving objects into the background. However, their results in a global framework are always below the ones obtained with supervised methods that use the groundtruth for training, as shown in the Results section of www.changedetection.net. Thus, in a second step, we use the output of this system as input to a refining convolutional neural network (CNN). It is based on a semantic segmentation network which is trained to accurately segment the objects that are present in an input mask.

Our contributions can be summarized as follows:

- We propose a new two-steps system for foreground segmentation which provides results superior to the ones obtained by conventional, non-learning based methods.
- We compare different off the shelf methods for background subtraction as input to the refinement CNN.
- We show that a Semantic Segmentation Network trained to refine the rough segmentation produced by a conventional background subtraction method can improve the results of background subtraction, without specific scene training.
- We provide an example of how learning based techniques can be used in conjunction with non-learning methods. This allows to introduce high level features according to the application and scenario, which currently cannot be used in most deep learning systems.

II. State of the Art

A. Classical approaches to foreground detection by background modeling and subtraction

Classical approaches model the background of the scene and then perform a comparison with the current frame for classifying each pixel as foreground or background.

Background is sometimes modelled as a single image computed as the temporal average or median over a set of frames

During the development of this work the first author was a visitor at TOSHIBA Cambridge Research Lab. This work has been carried out with the support of this lab and project TEC2016-75976-R, by the Ministerio de Economia, Industria y Competitividad and the European Regional Development Fund.

[1]. However, most methods use a more complex, probabilistic approach, capable of handling challenging situations such as illumination changes or slight variations due to outdoors conditions. Some examples for that are the Gaussian Mixture Model (GMM) [2], or the Kernel Density Estimation (KDE) [3]. Currently, the methods that obtain the best results for background modeling ([4], [5]) are based on robust median estimation and robust PCA, respectively. Joint background modeling and subtraction has been approached in many different ways. Stauffer and Grimson [2] proposed one of the first methods which is still widely used. This method models each pixel PDF as a mixture of Gaussians and uses an online approximation to update the model. By determining the Gaussian distributions that correspond to the background, each pixel can be labeled as foreground or background.

The great performance of that algorithm gave rise to a serial of different methods that applied slight variations for improving its weakest points [6], [7], [8].

However, not all classical approaches are parametric nor based on Gaussians. One example for these alternative methods is [9] that directly estimates the distribution for all background pixels instead of its parameters, thus performing in a non-parametric way and making use of the pixel's histogram distribution. Another non-parametric method for moving object detection is presented in [10]. It achieves high quality detections even in complex background scenarios and for non-completely static scenes, by dynamically estimating the bandwidth of the kernels used in the modeling part and selectively updating the background model. This method is improved in [11] by modeling not only the background, but also the foreground. At each new frame, for this approach, the spatial positions of the foreground data are updated using a particle filter that predicts its most probable movement.

B. Deep learning approaches to foreground detection

Attempts to obtain a good foreground/background pixel classification by applying deep learning techniques can be divided in those that focus on a successful background modeling and those that use this technique in the classification part. An extensive review can be found in [12].

1) Background modeling: Successful techniques for background modeling using deep learning include those based on Auto Encoder Networks [13] and on fully convolutional networks used for semantic segmentation such as U-Net, or Generative Adversarial Networks [14]. All these background modeling techniques rely on learning the background model for a specific scene, and thus cannot be easily updated to adapt to continuous changes in the background, incorporation of moving objects that remain static, etc. For this reason, although they provide excellent results with available databases, they are hard to apply to real scenarios conditions.

2) Background subtraction: Trying to improve traditional methods, many researchers have applied deep learning for classifying each pixel as foreground or background. Examples are [15], which is trained on the test scene or [16], which is trained with a subset of all the sequences of the database, which are used both for training and testing.

More recently, Generative Adversarial Networks have also been used. By considering the work of Phillip et al. [17] as a basis, *BScGAN* [18] is presented as a method for performing background subtraction and obtaining a foreground mask for each frame. The background model is a background image which is computed as the median of a set of images from the sequence. For the classification network, the structure of generator and discriminator from [17] is mostly preserved, but both original and background images are fed as input to the generator for obtaining the foreground mask. Although the obtained results are extraordinarily good, the method uses half of each sequence for training and tests it on the other half of the same sequence, thus lacking generalization and requiring annotations for most frames in a sequence.

III. Foreground Segmentation Network

Our proposed Foreground Segmentation Network is designed as a two steps procedure, with the objective to obtain a generic network that does not require learning from a specific scene. The first step detects the foreground objects according to the reasoning applied for the specific scenario, using classical background modeling techniques and an exception to model approach. This step provides a rough detection which is used as input to a refining network that is only trained once with a public domain labeled database, and that can be applied to any scenario without further adjustment. The global system thus combines the segmentation criteria and adaptability to background changes of the method used for the rough approximation with the characteristic power and capability of deep learning techniques for producing a much more precise segmentation.

A. Foreground detection

As a rough Foreground object detection a general method with a very low computational load has to be used. Although it can be substituted for any other more sophisticated detection module, we have preferred to use a common and widely available system in order to focus on the improvements of the refinement network. Three different methods provided by OpenCV have been considered. Although employing a different algorithm, they all use background models that are updated over time. The examples used show how specific application criteria can be introduced in this first stage. For instance, shadow suppression in the first one, or a high recall in the third. These methods also contain user-defined parameters which allow for a faster adaptation to background changes, the introduction in the background model of foreground objects that remain static or the degree of variability of the background depending on the scenario.

a) ***Mixture of Gaussians (MOG):*** This method, presented in [6], tackles the main downsides of the classical Stauffer and Grimson approach [2]: improve the learning at the beginning, especially in busy environments where the frames used for initializing the model contain a lot of foreground objects, and introduces shadow suppression.

*b) **Mixture of Gaussians2 (MOG2)**:* The second method considered [7] is also based on [2], but the update equations are adapted, and the number of Gaussian components for each pixel model is automatically selected.

*c) **Non Parametric Model (NPM)**:* This method ([9]) combines statistical background image estimation and per-pixel Bayesian segmentation. It is designed to maximize the F2-score, since for its specific application high recall is preferred over high precision.

B. Refinement network

Inspired by the works of semi-supervised Video Object Segmentation [19], our refinement network produces the foreground object segmentation guided by an input mask. In [19] a CNN designed for semantic image segmentation is used to produce label propagation in a video object segmentation context. For each new video frame the network is guided towards the object of interest by feeding in the previous frame mask estimate. The segmentation of the object in the first frame is manually provided. The network is trained with annotated images in order to produce a refined mask output for the current frame, given a rough mask estimate from the previous frame t-1. In order to produce accurate results, the network is fine-tuned on the manual annotation provided for the first frame and its augmentations.

In our work the refinement network takes as input the rough mask provided by one of the foreground detection methods described in Section III-A, which is concatenated to the current frame. A convolutional neural network designed for semantic segmentation is trained to refine this rough mask, given the current input image. Unlike [19], no training is performed on the specific sequence, thus no manual annotation is required at test time. In our case the guided semantic segmentation network is trained to classify pixels in two classes (foreground and background). The structure of the network is that of a U-Net. In particular, we have used the structure of the Generator in [17]. Details of the architecture can be found in this paper. In our case, the encoder is fed by 2 concatenated images resized to 256x256, corresponding to the current RGB image to segment and the rough binary mask approximation obtained by the Foreground detection block III-A. The output is a single image of the same size, corresponding to the refined mask. For training, we apply Data Augmentation to the input images to the Refinement network, which we detail in the following.

a) Data Augmentation: As it will be described in next Section, databases available for training are large in the number of frames, but reduced in the number of videos. As a consequence there is a high risk of overfitting to the training scenarios. To reduce this effect a large Data Augmentation is introduced, using the Albumentations package [20]. Each frame is converted using: horizontal flipping (50% probability), scaling (up to 0,5), rotation (20 degrees limit), shifting, random cropping to 256x256, gaussian noise addition, perspective transformation, blurring, brighthess and contrast transformation.

IV. Results

A. Datasets

Two datasets are used for the experiments: CDNet-2014 [21] and LASIESTA [22]. The first one is used for training the network and finding the optimal set of parameters, while the second, which is significantly smaller, is only used for testing and comparing with other State of the Art methods.

CDNet-2014 dataset contains a total of 53 video sequences, divided in 11 categories representing challenging scenarios such as dynamic background, hard shadows, low framerate or night videos, among others. Each of these categories contains from 4 to 6 sequences of 600 to 7999 frames with spatial resolutions varying from 320x240 to 720x576.

Annotation is provided for a number of frames for each sequence, which usually goes from 1/4 to 1/2 of the total number of frames. Training of the network and hyperparamenter search are carried out using these annotated frames.

In order to prove the applicability of the network in new datasets without further training or adaptation, a different dataset is used for testing. We use for this aim LASIESTA dataset. This dataset contains 10 different categories, 6 focused on indoor scenarios and the other 4 recorded in the outdoors. For each of these categories, two sequences with lengths ranging between 225 and 1400 frames are given. These sequences have a spatial resolution of 352x288. Each frame has its corresponding annotation, with the foreground moving objects.

We will use the standard evaluation metrics for assessing the performance of background subtraction algorithms: **Precision**, **Recall** and **F-measure**. While Precision indicates which amount of pixels detected as foreground are correct, recall gives a measure for the quantity of actual foreground pixels that are being correctly classified by the method, and F-measure gives a good general perspective of both criterion, thus being the one used for ranking methods and deciding the best one to use.

B. Experiments

In the first place, a train-validation partition of CDNet-2014 is used for the hyperparameter search experiments and for comparing different input masks. The selected hyperparameters and method of generation for input masks are used for training a refining network which can be used on a general scenario. We provide results for our CDNet validation partition and for LASIESTA database.

1) Network Hyperparameters: Hyperparameters are searched with a CDNet partition which is built by randomly selecting one sequence in each category for validation and the remaining ones for training. This gives 11 sequences in the first subset and 42 in the second, thus maximizing the amount of frames used for training but maintaining the diversity in the validation subset. The refinement network is trained taking as additional input the rough masks produced by MOG2, and Data augmentation is applied as described. The batch size used for training is 5. The Adam solver, with learning rate of 0.0001 provides best results. Multiple Cross Entropy is generally used for Semantic Segmentation. However, in our case, with only two classes, we have the option to use Binary Cross Entropy or Dice loss, which is more directly related to the F-measure that we are trying to optimize. The experiments performed show similar performances, although slightly better for the Dice loss. Convergence of the network occurs around 15 epochs, as shown in Figure 1. This Figure shows the F-measure results for 20 epochs, for Training and Validation data, using the parameters detailed and Dice loss.

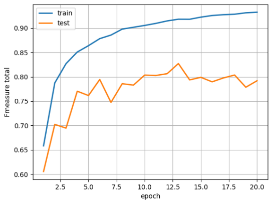

Fig. 1. F-measure evolution for training and validation partitions

2) Effect of Foreground detection: Three classical methods are considered for the Foreground detection block: NPM [9], MOG [6] and MOG2 [7]. Since all of these methods need some frames for initializing, the next variation is applied to each of them: First, they are initialized with frames in the first half of the sequence (a maximum of 120, 200 or 500, respectively) for extracting masks for the whole sequence. However, only the masks for the second half of frames are considered. Then, the sequence is reversed and the initialization is done on the second half for finally obtaining the masks for the remaining frames using the same reasoning. We emulate in this way the results that we could expect in a real scenario where the initialization frames can be discarded and we would only consider the results after a few minutes of initialization time.

The results of these methods for each category is first studied. The method that performs the best in the categories where the background is completely static, there is no color camouflage and neither the framerate of the sequence or the velocity of the objects change drastically, is NPM. This method was designed for a specific purpose with many constraints and performs well when they are fulfilled. However, the model is not prepared for dealing with more challenging situations, leading to a much lower performance in the other categories. Nonetheless, it should be noted that its goal, as stated in the paper, is to maximize F2-measure, meaning that high recall is prioritized over high precision, and it definitely is the model that obtains the highest recall in almost all of the categories. Similarly, the one with highest precision in almost all categories is the MOG method. Moreover, due to its high adaptability to changes in scene, it outperforms the other two methods also in F-measure for the categories in which the background is less static. Finally, thanks to the adaptive number of Gaussians used to model each pixel, MOG2 seems to obtain the best masks in those cases where the image drastically changes from one frame to another and also when the appearance of the foreground object is most similar to the background one. It is also the method with the highest F-measure considering all the sequences in the validation partition, as it can be observed in Table I, first row.

Next, the network is trained for 15 epochs with the previously selected hyperparameters, using generated masks from these methods as input. As it is shown in Table I, the refinement network increases the performance in all cases. Overall, MOG2 is the method that gets the better output masks. In the second row of this table we can compare the F-measure at the output of the Refinement Network, for each method for obtaining the input masks.

Method	NPM	MOG	MOG2
F-measure	0.35	0.50	0.53
F-measure Refined	0.64	0.75	**0.80**

TABLE I

3) Comparison to State of the Art methods: The experiments performed in the previous Section, confirm the improvement of the refinement network, and that best results can be achieved by selecting a rough mask detection method according to the scenario of interest. However, to assess the improvement in a general setting, the network trained with MOG2 input, which provided the best results, is used on a new test dataset (LASIESTA). This last set is made of 20 different video sequences corresponding to 10 different challenging scenarios. Some of these scenarios, like rainy or snowy conditions are similar to the ones in the training database. Other categories like occlusions or camouflage do not have an equivalent category in this database.

The F-measure obtained for each of the categories and the average for all sequences are reported in Table II. In this table, the performance for most methods that have been reportedly tested on this dataset are given, all run with a single set of parameters along the sequences and without any kind of training on this same dataset. Additionally, the metrics for the masks that are used as input to the mask refinement network are also provided, obtained by using the MOG2 algorithm [7]. Finally, the last column of the table correspond to the output of the method developed in this work. The evaluation method

	[23]	[2]	[8]	[10]	[24]	[11]	[7]	Ours
SI	0.82	0.83	**0.90**	0.78	0.88	0.88	0.67	0.88
CA	0.75	0.82	0.83	0.73	**0.89**	0.84	0.68	**0.89**
OC	0.88	0.88	**0.95**	0.85	0.92	0.78	0.72	0.92
IL	0.48	0.29	0.23	0.79	**0.84**	0.64	0.41	0.82
MB	0.74	0.76	0.86	0.72	0.84	**0.93**	0.68	0.78
BS	0.47	0.36	0.53	0.58	0.68	0.66	0.55	**0.85**
CL	0.85	0.86	0.87	0.91	0.82	**0.92**	0.56	0.80
RA	0.85	0.78	0.87	0.80	0.89	0.86	0.67	**0.90**
SN	0.59	0.60	0.38	0.45	0.17	0.77	0.38	**0,86**
SU	0.75	0.72	0.71	0.73	0.85	0.72	0.71	**0.87**
AVG	*0.72*	*0.69*	*0.71*	*0.73*	*0.78*	*0.80*	*0.60*	***0.86***

TABLE II

used is the one proposed in the database official webpage, in which only proper background pixels and foreground pixels corresponding to moving objects are considered.

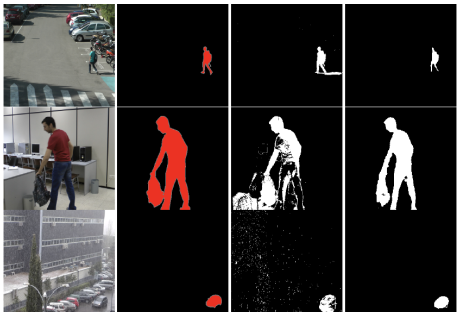

Fig. 2. Examples for the evaluation of the mask through the mask refinement method. First column: Input images. Second column: Corresponding annotation. Third column: Input mask obtained by MOG2. Forth column: Mask generated by the mask refinement network.

Analysing these results, it can be observed that the network has learnt how to properly generalize to unknown sequences even from a different nature than the ones used for training. The network has actually learnt how to refine the masks obtained with a classical method, increasing the performance in a 26%, reducing its noise, filling up holes in figures and discarding the background pixels that were mistakenly detected as foreground (Figure 2). Comparing to the other State of the Art methods, the presented mask refinement network seems to introduce a valuable approach. While it obtains a 86% of F-measure in all of the categories, it is the one with the best performance in several categories, and in the overall results.

V. Conclusions

This paper presents a novel approach for successfully subtracting the background from any sequence without requiring the annotation for any of its frames. Previous methods attempting this task with CNNs use a huge amount of annotations and specific training for each individual sequence. Our method uses a semantic segmentation network adapted to classify into two classes, foreground and background. It is composed of two steps: the first one uses a conventional foreground detection method to provide a rough approximation to the masks. Then, the CNN uses this approximation as a guidance and outputs a refined foreground mask.

We compare different classical methods for obtaining the inputs of the network. Using the best method and the optimal set of parameters, the network is trained on one dataset (CDNet-2014) and tested in another one (LASIESTA), obtaining satisfactory results. The network obtained captures more detail in the foreground masks than the state of the art methods and detects more compact foreground components.

References

[1] B. Laugraud, S. Piérard, M. Braham, and M. Droogenbroeck, "Simple median-based method for stationary background generation using background subtraction algorithms," 08 2015.

[2] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *Proceedings of IEEE Conf. Computer Vision Patt. Recog*, vol. 2, 01 1999.

[3] A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *Proceedings of the 6th European Conference on Computer Vision*, ser. ECCV, 2000, pp. 751–767.

[4] B. Laugraud, S. Piérard, and M. Droogenbroeck, "Labgen-p: A pixel-level stationary background generation method based on labgen," 12 2016.

[5] S. Javed, A. Mahmood, T. Bouwmans, and S. Jung, "Background-foreground modeling based on spatiotemporal sparse subspace clustering," *IEEE Transactions on Image Processing*, vol. PP, 08 2017.

[6] P. Kaewtrakulpong and R. Bowden, "An improved adaptive background mixture model for realtime tracking with shadow detection," *Proceedings of 2nd European Workshop on Advanced Video-Based Surveillance Systems; September 4, 2001; London, U.K*, 05 2002.

[7] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition*, 2004, pp. 28–31 Vol.2.

[8] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recogn. Lett.*, vol. 27, no. 7, pp. 773–780, May 2006.

[9] A. B. Godbehere, A. Matsukawa, and K. Goldberg, "Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation," in *2012 American Control Conference (ACC)*, June 2012, pp. 4305–4312.

[10] C. Cuevas and N. García, "Improved background modeling for real-time spatio-temporal non-parametric moving object detection strategies," *Image and Vision Computing*, vol. 31, p. 616–630, 09 2013.

[11] D. Berjón, C. Cuevas, F. Morán, and N. García, "Real-time nonparametric background subtraction with tracking-based foreground update," *Pattern Recognition*, vol. 74, 09 2017.

[12] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *CoRR*, vol. abs/1811.05255, 2018.

[13] P. Xu, M. Ye, X. Li, Q. Liu, Y. Yang, and J. Ding, "Dynamic background learning through deep auto-encoder networks," in *Proc of the 22nd ACM International Conf on Multimedia*. ACM, 2014, pp. 107–116.

[14] M. Sultana, A. Mahmood, S. Javed, and S. K. Jung, "Unsupervised deep context prediction for background foreground separation," *CoRR*, vol. abs/1805.07903, 2018.

[15] M. Braham and M. Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," 05 2016, pp. 1–4.

[16] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognition*, vol. 76, pp. 635–649, 2018.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of IEEE CVPR*, 2017, pp. 1125–1134.

[18] M. C. Bakkay, H. A. Rashwan, H. Salmane, L. Khoudour, D. Puigtt, and Y. Ruichek, "BSCGAN: deep background subtraction with conditional generative adversarial networks," in *2018 IEEE International Conference on Image Processing, ICIP*, 2018, pp. 4018–4022.

[19] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *Proceedings of IEEE CVPR*, 2017, pp. 2663–2672.

[20] E. K. V. I. I. A. Buslaev, A. Parinov and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *ArXiv e-prints*, 2018.

[21] Y. Wang, P. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "Cdnet 2014: An expanded change detection benchmark dataset," in *IEEE CVPR Workshops*, June 2014, pp. 393–400.

[22] C. Cuevas, E. María Yáñez, and N. García, "Labeled dataset for integral evaluation of moving object detection algorithms: Lasiesta," *Computer Vision and Image Understanding*, vol. 152, 08 2016.

[23] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, July 1997.

[24] T. S. F. Haines and T. Xiang, "Background subtraction with dirichlet-process mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 670–683, April 2014.