An Efficient Pixel-wise Method for Moving Object Detection in Complex Scenes

Weiguo Feng*, Rui Liu*, Baozhi Jia†, and Ming Zhu†

*Department of Electronic Engineering and Information Science, University of Science and Technology of China
{fwg168, liuruin}@mail.ustc.edu.cn
†Department of Automation, University of Science and Technology of China
bobby32@mail.ustc.edu.cn, mzhu@ustc.edu.cn

Abstract

Moving object detection is often one of the most basic and important stages in computer vision applications. In this paper, a novel background model is proposed to extract moving foreground objects from videos that may contain different kinds of disturbance such as illumination changes, camera parameter variations, noises and dynamic backgrounds, etc. For each frame, a local frequency response map is generated using short-term Fourier transformation (STFT) in local regions, and by extracting the relations among neighborhoods of the response map, a compact pixel feature is introduced as local frequency pattern. Then, an adaptive probabilistic estimation of pixel feature sequence modified from kernel density estimation is performed to estimate the probability of a pixel being background. Experimental evaluations on complex scenes of surveillance videos demonstrate that the proposed method has archived satisfactory results.

1. Introduction

With massive of surveillance cameras deployed in the public world, automated analysis and processing of surveillance videos are crucial, as visually combing in such huge video data sets is impractical for human analysts. Extracting moving foreground objects from monitored videos is a critical procedure in the smart surveillance video processing applications. The extracted foreground objects are usually used as input to further high-level process, such as object tracking, event alert and behavior analysis, etc. Thus, its performance can have a huge influence on the further processing. However, due to various complex scenes in the real world, moving foreground object detection is still a challenging task and a hot research topic in the past few decades. One of the most popular approaches presented in the literature is to construct a background model that classify each frame pixel being background or not according to its feature values, which is referred to as background subtraction. Some excellent surveys on the recent development of these methods can be found in [1, 2].

The most widely used method is Gaussian Mixture Model (GMM) proposed by Stauffer and Grimson [3], in which the pixel process is modeled by a mixture of K weighted Gaussian distributions in color space. Different from a fixed number of Gaussian distributions used in the original GMM, Zivkovic [4] and Lee [5] presented a modification by adaptively choosing the number of mixture components for each pixel. GMM-like methods have been widely used in different forms because of their simplicity and computational efficiency. However, their methods cannot handle the disturbance of sudden intensity variations and highly dynamic backgrounds, yielding lots of false detections. Considering that the assumption of pixel processes subjecting to Gaussian distribution is restrict, Elgammal *et al.* [6] took advantage of kernel density estimation to provide a more flexible model. By estimating the probability of observing pixel based on samples of intensity values of each history pixel in a non-parametric way, their method can adapt quickly to the dynamic changes in the scenes. Kim *et al.* [7] proposed a compact non-parametric algorithm for foreground detection, which sampled background pixel values and quantized them into a set of codewords.

More recently, texture features have been adopted by lots of computer vision applications and are also introduced to extract features for moving object detection. Heikkilä *et al.* [8] presented a block-based method which employed textures features by modeling each block with its local binary pattern (LBP) histogram to capture the background statistics. Although the block-based methods are generally more

978-1-4799-0703-8/13/$31.00 ©2013 IEEE

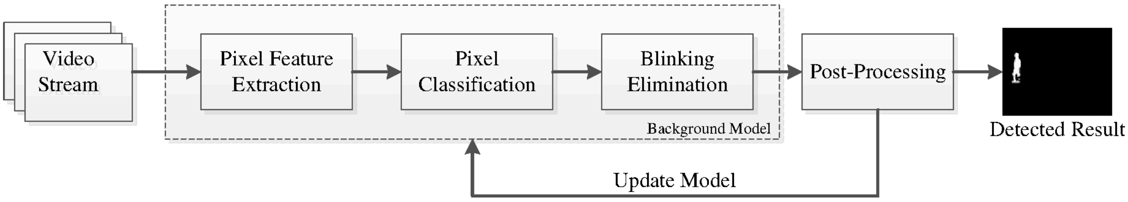

Figure 1. Flowchart of proposed moving object detection method.

robust and computational efficient than other pixel-based methods, the main drawback of these methods is that they can't extract the accurate shape of the moving object very well. Liao *et al.* [9] developed a scale invariant local ternary pattern (SILTP) operator for background model together with the pattern kernel density estimation for probability estimation. Different from the previous traditional methods, Zhou and Xu *et al.* [10] modeled the moving foreground object detection as an image labeling problem, in which both spatial and temporal coherency of the object were enforced using an Markov Random Field (MRF) model. Thus, the final detection mask is determined by inferring the maximum a posterior. However, in the these MRF-based methods, it's quite critical to design the appropriate components of the energy function.

In this paper, we propose a novel and robust approach for tackling the moving object detection from surveillance videos containing different kinds of disturbances. For each frame, a local frequency response map is generated using short-term Fourier transformation (STFT) in local regions, and each pixel is further encoded by the relations among the neighborhood of the response map. Then, the pixel feature sequence is modeled by an incremental adaptive estimation modified from kernel density estimation. Finally, the new coming pixel is evaluated belonging to background by comparing its distribution probability with a given threshold. Extensive experiments have been carried out and the results show that the proposed method is very efficient and robust under complex video scenarios such as sudden changes of illumination and rippling water surfaces.

The rest of this paper is organized as follows. Section 2 describes the detail of the proposed method, including the pixel feature extraction, background modeling of the feature process and blinking elimination. Section 3 presents a comprehensive comparison of our method with other methods, followed by the conclusions in Section 4.

2. Proposed Method

The approach proposed for moving object detection in this paper follows the processing flow as shown in Fig. 1. For each frame, a compact pixel feature referred to as local frequency pattern (LFP) is extracted by the relation of local frequency response among its neighborhoods. With a history process of such pixel feature sequences, a probabilistic model is trained to estimate the probability of a pixel being background. Furthermore, a heuristic method is applied to eliminate the blinking pixel in the detected mask. The details of each components in the method will be described in the following subsections.

2.1. Local Frequency Pixel Pattern

Traditionally, pixel intensities and color values are often used as feature description method in background model [3, 4, 6]. And with great success in facial recognition and image classification, spatial features such as edges, gradients and texture features, etc. are also introduced to moving object detection [8, 11]. However, these methods use spatial information to encode pixels, which are very vulnerable to image noises and highly dynamic backgrounds. In frequency analysis, noises, waving trees, drifting clouds, and rippling water surfaces of the complex backgrounds often contribute plenty to the high-frequency part of the image. Therefore, in this work we introduce a local frequency pattern operator, in which only low frequency part of the image is extracted to encode each pixel.

Given an input image $f(\mathbf{x})$, its local frequency information is extracted using a short-term Fourier transformation computed over a rectangular $M \times M$ neighborhood $\mathcal{N}_{M\times M}(\mathbf{x})$ at each pixel position $\mathbf{x}$, which is defined by

$$F(\mathbf{u},\mathbf{x}) = \sum_{\mathbf{y}\in\mathcal{N}_{M\times M}(\mathbf{x})} f(\mathbf{x}-\mathbf{y})\exp\{-j2\pi\mathbf{u}^T\mathbf{y}\} \quad (1)$$

where $\mathbf{u} = [u, v]$ is the frequency vector in horizonal and vertical directions. In the real world applications, real-time processing is crucial for moving object detection. As can be noticed that, Eq. (1) can be derived to

$$F(\mathbf{u},\mathbf{x}) = \mathbf{w}_{\mathbf{u}}^T\mathbf{f}_{\mathbf{x}} \quad (2)$$

where $\mathbf{w}_{\mathbf{u}}$ is the vectorized convolution kernel at frequency $\mathbf{u}$, and $\mathbf{f}_{\mathbf{x}}$ is the pixel vector containing all M^2 pixel samples from sub-image $\mathcal{N}_{M\times M}(\mathbf{x})$. Since the convolution kernel is separable, the final transform can be computed using 1-D convolution for rows and columns successively. Fig. 2 shows local frequency response maps by performing STFT in different frequency.

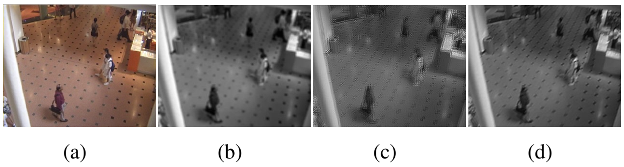

Figure 2. An example of transforming frame (a) using STFT in different frequency vector $\mathbf{u}$: (b)$\mathbf{u} = [1, 1]$, (c)$\mathbf{u} = [1/3, 1/3]$, (d)$\mathbf{u} = [1/9, 1/9]$.

The local frequencies extracted from low-frequency part of the image can capture the local texture properties and filter out high-frequency part caused by noises, highly dynamic backgrounds, etc. This imposes advantages over the intensity and color based features used in the traditional approaches. Particularly, in order to provide more robustness, we adopt the idea of local binary pattern (LBP) to quantize the relation pattern among neighborhoods [8]. In LBP, the image pixels are labeled by thresholding the neighborhoods of each pixel with the center value and the result is considered as a binary number, which is referred to as the new pattern. In the local frequency response map, we denote the magnitude of the response at position $\mathbf{x}$ and frequency $\mathbf{u}$ as $M(\mathbf{x})$, and the 3×3 neighborhood centered on $\mathbf{x}$ as $M(\mathbf{y})$ where $\mathbf{y} \in \mathcal{N}_{3\times3}(\mathbf{x})$. The relation between the neighborhood and the center is quantized by sign function $s(\cdot)$

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{3}$$

The local frequency pattern at position $\mathbf{x}$ and frequency $\mathbf{u}$ can be computed by concatenating the eight bits into a scalar value

$$LFP_{\mathbf{u}}(\mathbf{x}) = \sum_{i=0}^{8} s(M(\mathbf{y}_i) - M(\mathbf{x})) \times 2^i \tag{4}$$

where $\mathbf{y}_i$ is the i-th neighbored value of $\mathcal{N}_{3\times3}(\mathbf{x})$. Thus the new pattern of the pixel is in the range 0-255. Fig. 3 presents a illustration of LBP quantization of local frequency response map.

With the advanced properties of local frequency analysis and local binary pattern operator, the local frequency pattern can model the pixel information in frequency domain.

2.2. Background Model Estimation

The goal of every moving object detection method is to classify each pixel of the image into two categories, background pixel and moving foreground pixel. In this subsection, an incremental adaptive probabilistic model is introduced to estimate the probability of a pixel being a moving foreground pixel [6]. Suppose a feature sequence is derived from local frequency pattern operated on a set of t history background samples and is defined as

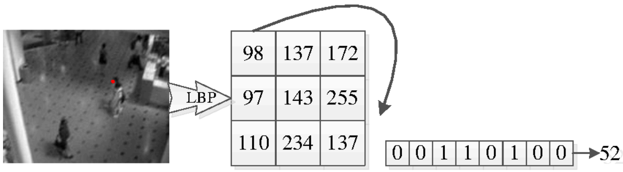

Figure 3. Illustration of performing LBP quantization of the local frequency response.

$$\{p^0, \ldots, p^{t-1}\} = \{LFP^0_{\mathbf{u}}(\mathbf{x}), \ldots, LFP^{t-1}_{\mathbf{u}}(\mathbf{x})\} \tag{5}$$

Under a KDE model, the new coming pixel is determined belonging to background or foreground according to its history feature sequence. The probability of the pixel under the background model can be estimated by

$$Pr(p^t|\mathcal{B}) = \sum_{i=0}^{t-1} \omega_i K_H(p^t - p^i) \tag{6}$$

where ω_i is the weighting coefficient of each history sample and K_H is the kernel estimator with bandwidth H which is a symmetric function that integrates to one. There are several kernel estimators commonly used in KDE: Epanechnikov kernel, uniform kernel, Gaussian kernel, and others. In our work, Gaussian kernel is adopted due to its convenient mathematical properties. Note that choosing Gaussian function as the kernel estimator is different from fitting the pixel feature process to Gaussian distribution. The research also has shown that choosing which kernel function does not affect too much on the final estimate, whereas the bandwidth of the kernel is a crucial parameter which exhibits a strong influence of the estimation. And in the real-time applications, only a limited length of pixel feature process can be stored for the modeling. In order to achieve an efficient and accurate estimation result, the kernel bandwidth is trained from successive absolute deviation of the pixel feature sequence [12]

$$\Delta_p = \{|p^i - p^{i-1}| \quad | \quad i = 1, 2, \ldots, t-1\} \tag{7}$$

The covariance of the successive absolute deviation can represent the temporal scatter of the training samples, which makes itself a good choice for kernel bandwidth

$$\sigma^2 = cov(\Delta_p) = (\Delta_p - \mu_\Delta)(\Delta_p - \mu_\Delta)^T \tag{8}$$

where μ_Δ is a length-$(n-1)$ vector with each element equal to the mean value of Δ_p. Thus, with the adaptive chosen kernel bandwidth, the final probability of a pixel being background can be estimated as

$$Pr(p^t|\mathcal{B}) = \sum_{i=0}^{t-1} \frac{\omega_i}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(p^t-p^i)^2}{\sigma^2}\right\} \tag{9}$$

After the probability of each pixel feature in the new frame is estimated, the median of the probabilities in 8-connected neighborhood of each pixel is compared with a given threshold to determine whether the pixel belongs to background

$$M(\mathbf{x}) = \begin{cases} 0, & \text{if } median(Pr(p^t|\mathcal{B})) \geq thr \\ 1, & \text{otherwise} \end{cases} \tag{10}$$

2.3. Background Model Update

Once the background model is constructed, updating the model is essential for tackling the variations in the background, such as illumination changes, shadows and dynamic backgrounds.

Since the foreground regions have nothing to do with the background model, only pixels classified as background are used for updating in our method. In the surveillance videos, due to the exists of dynamic background objects, the pixels belonging to those regions may often misclassified as foreground pixels. To overcome these variations, we adopt an update mechanism that searches the local region of each pixel in the background model for the best match. The best match is defined to be the neighboring sample with minimum distribution probability of the current pixel model

$$\mathbf{y}_m = \arg\min_{\mathbf{y}\in\mathcal{N}_\mathbf{x}} Pr(p^t_\mathbf{y}|\mathcal{B}) \tag{11}$$

Considering a pixel of the waving tree branches, it will move to its neighborhood as time goes, and then move back to its original position. In [13], a similar approach called "diffusion" is used to update the background model, which is based on a random sampling scheme. Instead, our approach chooses the neighboring pixel based on the likelihood between the sample feature and the background model, which provide more robustness against choosing the pixel randomly. The new match is then used to update the background model

$$Pr^t(p^t|\mathcal{B}) = (1-\rho)Pr^{t-1}(p^t_{\mathbf{y}_m}|\mathcal{B}) + \rho K_H(p^t_{\mathbf{y}_m}, p^{t-1}) \tag{12}$$

where $Pr^t(p^t|\mathcal{B})$ denotes the distribution probability of background model at time t and $p^t_{\mathbf{y}_m}$ is the feature of the found pixel match $\mathbf{y}_m$.

2.4. Blinking Pixel Elimination

In the moving object detection applications, complicated scenes and different kinds of variations make it impossible to present an explicit model to adapt to them, which leads to lots of misclassification and false detection. Especially, those pixels that often switches between background and foreground contribute much to the background model estimation and make the model unstable for these variations. In this subsection, we adopt a processing trick proposed in [14] to eliminate these blinking pixels.

The idea of dealing with the blinking pixel is to check if there are too many switches between background and foreground in a short classification history sequence. Thus, a map with blinking counts is generated according to the previous updating mask for each pixel. If the current updating mask of a pixel is different from its previous updating label, the blinking counts is increased by 15, otherwise the counts is decreased by 1. When the blinking counts reaches 0, it can not be reduced anymore.

A pixel is determined to be blinking if its blinking counts is larger than 50. Then the pixel is removed from the updating mask and is classified as background. This trick can enhance the robustness of our algorithm for sample based models and can provide a more compact detection result.

3. Experimental Results

The proposed method has been evaluated exhaustively on complex I2R dataset[1], which contains a variety of indoor and outdoor environments, including rippling water surfaces, waving tree branches, etc. The I2R dataset consists of nine video sequences captured in diverse and challenging environments, grouped in two catagories: six indoor scenes(Hall, Bootstrap, Curtain, Escalator, Lobby and ShoppingMall), and three outdoor scenes(Fountain, Trees and WaterSurface). For each scene in the dataset, there are over several thousands video frames and 20 randomly selected frames that are labeled manually as groundtruth.

In the experiments, the proposed method was compared with the state-of-the-art GMM [3], ACMMM 03 [4], block-based LBP histogram approach [8], codebook approach [7] and SILTP method [9]. In the qualitative comparison, for the sake of fairness, no morphological operations or any other postprocessing methods were applied to the resulting foregrounds for all the algorithms. And a set of consistent parameters of our approach were chosen as: $\mathbf{u} = [1/9, 1/9], M = 5, \rho = 0.003, thr = 0.75$. Due to the limitation of the paper length, three typical videos in the dataset were selected for analysing the results. Fig. 4 shows qualitative results on several kinds of videos sequences of GMM, ACMMM03, CodeBook with default

[1] http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html

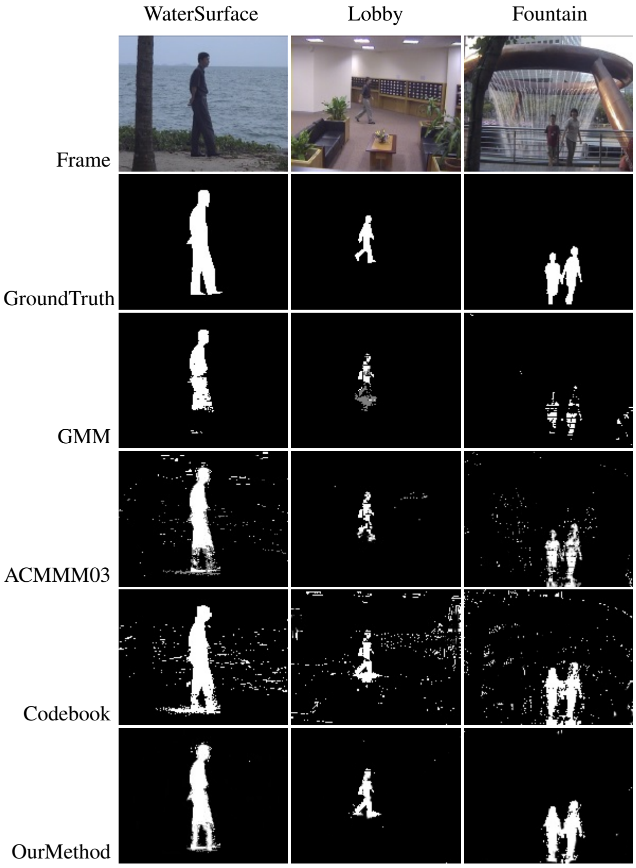

Figure 4. Visual comparison of the detected moving foregrounds on different video sequences.

parameters. As demonstrated, the color-based GMM suffers from false detections caused by moving shadows. Although ACMMM03 removes the soft shadows successfully, many small foreground regions are missed in the detection results. In codebook approach, there are many snowflake pixels which may yield lots of difficulties for highlevel applications. By extracting local frequency information of each pixel, our method is more robust to illumination variance. And it can also tolerate small moving background regions by modeling the background with neighborhood spatial searching scheme.

In addition, a quantitative performance evaluation of the detection results is applied for all of the nine datasets with F-measure. The F-measure evaluates the detection accuracy by

$$F = 2 \cdot \frac{RC \cdot PR}{RC + PR} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \qquad (13)$$

where $F \in [0, 1]$ with RC, PR, TP, FP and FN being recall, precision, true positives, false positives and false negatives, respectively. The higher the F value, the more accurate the foreground subtraction. Tab. (1) illustrates a more quantitative performance results for our method against compared algorithms, in which the results of LBP Histogram and SILTP are reported by Liao *et al.* [9].

4. Conclusion

Moving object detection is often one of the most basic and important stages in the surveillance video applications. In this paper, we proposed a novel pixel-wise background modeling method for moving object detection in scenes with different kinds of complex disturbance, such as illumination variations, camera exposure changes and dynamic backgrounds, etc. First, our method extracts frequency information from the local regions of the image using short-term fourier transformation (STFT), and the feature of each pixel is constructed by describing the relations among the neighborhoods of the local frequency magnitude. Then, an incremental adaptive probabilistic estimation extended from kernel density estimation is performed to evaluate the probability of a pixel being background. Finally, by adopting an elimination mechanism of blinking pixels, our method is more robust to complex scenes and dynamic backgrounds. Extensive experiments show that the proposed approach produces a reliable and accurate detection results on complex surveillance videos.

Acknowledgement

This research is supported by the National Key Technology R&D Program of China (Grant No. 2011BAH11B01) and the "Strategic Priority Research Program - Network Video Communication and Control" of the Chinese Academy of Science (Grant No. XDA06030900).

References

[1] M. Piccardi, "Background subtraction techniques: a review," in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 4, pp. 3099–3104 vol.4, 2004.

[2] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *Image Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 294–307, 2005.

[3] C. Stauffer and W. E. L. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," in *Computer Vision and Pattern Recognition*, vol. 2, pp. 2246–2252, 1999.

[4] Z. Zivkovic, "Improved Adaptive Gaussian Mixture Model for Background Subtraction," in *International Conference on Pattern Recognition*, vol. 2, pp. 28–31, 2004.

[5] D. shyang Lee, "Effective Gaussian Mixture Learning for Video Background Subtraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 827–832, 2005.

Videos	GMM	ACMMM03	LBPH	CodeBook	SILTP	OurWork
Bootstrap	53.78	60.25	52.81	56.89	75.35	67.43
Campus	42.47	78.62	62.85	65.73	42.54	79.38
Curtain	50.04	55.46	66.08	47.04	91.16	89.49
Escalator	31.64	29.01	59.08	48.93	63.90	65.14
Fountain	56.59	58.37	70.52	81.32	83.45	84.05
Hall	57.98	52.13	47.73	53.40	68.14	71.72
Lobby	65.43	71.35	50.29	69.81	78.80	80.47
ShoppingMall	67.96	65.29	54.67	57.95	79.62	79.29
WaterSurface	87.23	63.59	76.80	65.66	75.35	90.51

Table 1. Quantitative evaluation of foreground subtraction on the I2R dataset

[6] A. Elgammal, D. Harwood, and L. S. Davis, “Nonparametric background model for background subtraction,” in *European Conference on Computer Vision*, pp. 751–767, 2000.

[7] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, “Real-time foreground–background segmentation using codebook model,” *Real-time imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[8] M. Heikkila and M. Pietikainen, “A texture-based method for modeling the background and detecting moving objects,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 4, pp. 657–662, 2006.

[9] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikainen, and S. Z. Li, “Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 1301–1306, IEEE, 2010.

[10] Y. Zhou, W. Xu, H. Tao, and Y. Gong, “Background segmentation using spatial-temporal multi-resolution mrf,” in *Application of Computer Vision, 2005. WACV/MOTIONS ’05 Volume 1. Seventh IEEE Workshops on*, vol. 2, pp. 8–13, 2005.

[11] A. Ramirez-Rivera, M. Murshed, and O. Chae, “Object detection through edge behavior modeling,” in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pp. 273–278, 2011.

[12] A. Tavakkoli, M. Nicolescu, G. Bebis, and M. Nicolescu, “Non-parametric statistical background modeling for efficient foreground region detection,” *Machine Vision and Applications*, vol. 20, no. 6, pp. 395–409, 2009.

[13] O. Barnich and M. Van Droogenbroeck, “Vibe: A universal background subtraction algorithm for video sequences,” *Image Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1709–1724, 2011.

[14] M. Van Droogenbroeck and O. Paquot, “Background subtraction: Experiments and improvements for vibe,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 32–37, 2012.