A ROBUST MOTION DETECTION ALGORITHM ON NOISY VIDEOS

Yu Liu, Huaxin Xiao, Wei Wang, Maojun Zhang

National University of Defense Technology

ABSTRACT

The applicability and performance of motion detection methods dramatically degrade with the increasing noise. In this paper, we propose a robust dictionary-based background subtraction approach, which formulates background modeling as a linear and sparse combination of atoms in a pre-learned dictionary. Motion detection is then implemented to compare the difference between sparse representations of the current frame and the background model. The projection of noise over the dictionary being irregular and random guarantees the adaptability of our approach. Experimental results on synthetic and real noisy videos demonstrate the robustness of the proposed approach compared to other methods.

Index Terms—motion detection, dictionary learning, noise, sparse representation.

1. INTRODUCTION

Motion detection is defined as the problem of segmenting moving objects from a given image sequence or surveillance video. It has drawn considerable attention in the field of computer vision and video processing over the past decades. The most prevalent approach is background subtraction, which establishes a background model through a certain method and then calculates the difference between the current frame and the background to segment the foreground area. The statistical background model has been adequately researched and developed in past years. In [1]-[5], the mixture of Gaussian (MoG) model was established gradually and has been demonstrated to be a simple and effective approach. Oliver et al. [6] considered spatial configurations that captured eigen-backgrounds by eigenvalue decomposition based on the whole image, in contrast to methods using a statistical background model. Cevher et al. [7] assumed that most pixels in a frame belong to the background and introduced the theory of compressive sensing to achieve detection. Later, similar to the work presented by [6], incremental principal component analysis (PCA) [8] was used to capture the motion characteristics of backgrounds. A Self-Organizing approach to Background Subtraction (SOBS) [9] was proposed through artificial neural networks and achieved robust detection for different types of videos taken with stationary cameras. Lately, a visual background extractor (ViBe) [10] based on classification model can apply only one frame to update the background model and is one of the mainstream approach.

The aforementioned methods are focused more on handling complex and dynamic scenes, such as rain, snow, waves, and shaking trees, without considering the quality of images. Surprisingly, within this well-understood area, an elementary problem regarding corrupted signals is commonly observed in practice but has not been explored in depth. In real applications, image signals can easily get polluted in many cases: such as low light surveillance, heat in the sensor or turbulence in the signal transmission, where high level of noise causes existing algorithms to perform inappropriately. One approach to solve this problem is to employ denoising methods before applying detection. However, even state-of-the-art denoising algorithms cannot guarantee that the image quality is adequate for detection because most detection algorithms assume purely unpolluted images.

In order to deal with this problem, we propose a motion detection algorithm based on the theory of sparse representation that is stable on noisy videos. The proposed method employs a dictionary learning algorithm to obtain bases and formulates a background modeling step as a sparse representation problem. It establishes the dictionary from spatio-temporal image patches and then projects the current frame on this trained dictionary to obtain a corresponding coefficient for its representation. Different scene contents ought to have different coefficients. That is, the foreground would not lie on the same subspace spanned by the background, helping us to identify changes in the scene by comparing spanned coefficients. Given that statistical noise is typically distributed through the entire space anisotropically, its influence on real signal will be weakened obviously after the process of sparse representation. This character enhances the robustness of the proposed method to corrupted signals, while retaining its capability of dealing with non-noisy images and dynamic scenes.

2. PROPOSED METHOD

As described in Section 1, the proposed method can be divided into three parts: background modeling, dictionary learning for arbitrary scenes, and two-stage foreground detection based on sparse coefficients. The flow chart of our method is illustrated in Fig. 1.

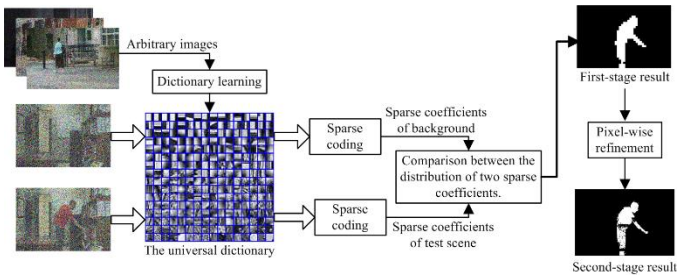


Fig. 1: The flow chart of the proposed method.

2.1. Background model

The background subtraction problem is usually formulated as the linear combination of a background model I_B and a foreground candidate I_F :

$$I = I_B + I_F \tag{1}$$

This study uses a pre-learned dictionary D to sparsely represent the background model, and the dictionary D is more general for arbitrary scenes. The background model is shown as follows:

$$I_B = D \times A \tag{2}$$

where A is the matrix of the sparse coefficients. The atoms d in dictionary D represent the bases of image signals. Via sparse coding, the background model can be regarded as a sparse and linear combination of the atoms.

2.2. Dictionary for arbitrary scenes

Compared to traditional methods of obtaining bases, such as wavelet and PCA, dictionary learning does not emphasize the orthogonality of bases, making its representation of the signal have better adaptability and flexibility. Dictionaries are effective for signal reconstruction and classification in the audio and image processing domain [11]. For a training set $Y = \{y_i\}_{i=1}^{N}$, its dictionary $D \in \mathbb{R}^{n \times k}$ satisfies the following formula:

$$D = \arg\min_{D} \sum_{i=1}^{N} \min_{\alpha_i} \left(\|y_i - D\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \right) \tag{3}$$

where k is the number of atoms in D, α_i are the sparse coefficients and λ is the regularization parameter.

In our approach, Online Dictionary Learning [11] is employed to cope with Equation (3). This algorithm exploits stochastic online dictionary learning adapted to sparse coding tasks which we use in the foreground detection step. Compared to the K-SVD [12], Online Dictionary Learning [11] runs faster and demands less memory. This study extracts arbitrary images from different scenes as the training set.

The size of the dictionary can be considered as a tradeoff issue. A larger scale of the dictionary can capture more detailed structures with longer time consumption in the process of dictionary learning and sparse coding. By contrast, a smaller dictionary performs worse in sparse representation but responds faster in detection. In this work, a relatively larger dictionary (256 atoms for the first stage and 81 atoms for the second stage) is used as the preferable option to achieve better detection performance.

2.3. Two-stage foreground detection

Following Equation (1), the foreground I_F is the difference between the current frame I and the background model I_B :

$$I_F = I - I_B = I - D \times A \tag{4}$$

However, adopting the corrupted frame I can still be directly affected by the noise. In this study, we project the current frame I over the learned dictionary and compute the sparse codes A' with LARS algorithm [13]. Then, the formula is converted as follow:

$$I_F = D \times A' - D \times A \tag{5}$$

We can then compare each patch to decide whether it belongs to the foreground through the distribution of the sparse codes:

$$\begin{cases} \Delta_1 = \left| \|\alpha_i\|_0 - \|\alpha_i'\|_0 \right| \\ \Delta_2 = \|\alpha_i - \alpha_i'\|_1 \end{cases} \tag{6}$$

where α_i and α_i' are the sparse coefficient of the ith block in the background model and the current frame. Δ_1 and Δ_2 are the differences of sparse coefficient's distributions and values between the background model and the current frame. Given that the distributions and values reflect which subspace is expanded by the test frame, we can use these parameters to decide whether or not the content of the monitoring scene has any movements. Specifically, if the image content remains the same, it tends to have identical distributions and corresponding values. By contrast, if a foreground object enters the scene and changes the content, it generates distinct distributions.

To obtain a more precise result, we post-process the differences of the sparse coefficients as follows:

$$\Delta'(i) = \gamma_1 \Delta_1(i) + \gamma_2 \Delta_2(i) \tag{7}$$

and

$$\Delta(i) = (1 - SST_i) * (\Delta'(i) + \sum_{k \in neighbor(i)} \Delta'(k)) \tag{8}$$

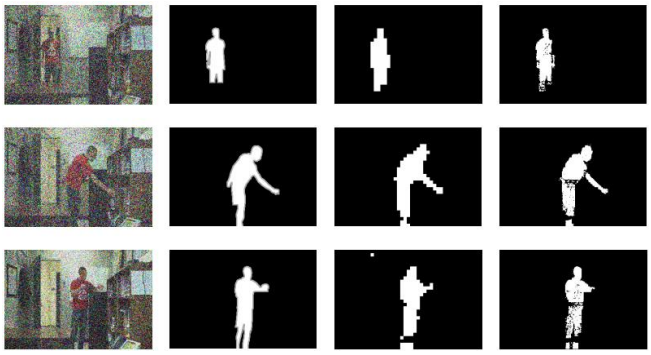
where i is the number of the image block. γ_1 and γ_2 are unitary parameters which determine the weights of Δ_1 and Δ_2 respectively. $neighbor(i)$ means the neighborhood blocks of the ith block. The Structural Similarity Testing (SST) is defined as follow:

$$SST(B_1, B_2) = \frac{(2\mu_{B_1}\mu_{B_2} + C_1)(2\sigma_{B_1B_2} + C_2)}{(\mu_{B_1}^{\ 2} + \mu_{B_2}^{\ 2} + C_1)(\sigma_{B_1}^{\ 2} + \sigma_{B_2}^{\ 2} + C_2)} \tag{9}$$

where B_1 and B_2 are the image blocks of I_B and $D \times A'$. μ and σ represent the average and variance of the image block. C_1 and C_2 are the parameters similar to the ones defined in [14]. SST is a weight parameter that represents

the structure similarity between the sparse representation of the background model and the current frame.

Equations (7), (8) and (9) can enhance the effect of the segmentation because the *i*th block belonging to a foreground object not only incorporates its own intensity but also of its neighborhood. If the sparse representation of the current frame block has similar structure as the one in the background model, the value of SST_i would be very large. In this case $\Delta(i)$ is low and the block would not be regarded as the foreground. Thus, the *SST* further improves the precision of the detection results.



(a)Noisy frames (b)Ground truth (c)1st stage (d)2nd stage

Fig. 2: (a): The three frames are extracted from **Office** dataset in [14] (Frame No. 611, 677 and 794) and are added the mixture noise of Gaussian and Poisson. (b): Ground truth. (c)-(d): First and second stage detection results with proposed method.

The dictionary based on patches leads to unsatisfactory precision, as shown in the Fig. 2(c). To obtain a pixel level result, the proposed method includes two-stage foreground detection. The first stage roughly detects the foreground in the patch, and the second stage refines the detection results on top of the first stage. Assume that the size of the atom in the first stage dictionary is $K \times K$. Thus, the size of the foreground candidate block is $K \times K$. For each foreground block, we use a $L \times L$ sliding window to determine whether the central pixel belongs to the foreground. Pixel-wise refinement as shown in Fig. 2(d) achieves more precise detection results.

Background updating. An important characteristic for any background subtraction algorithm is to continuously update the learned model over time. The update process is the ability to handle gradually changing illumination and adapt to new objects that appear in a scene. Since the dictionary in our work is learnt as preprocessing step by arbitrary images, the background update process is to update the sparse coefficients every frame or couple of frames according to the implementation requirements.

Parameter selection. We use the regularization parameter $\lambda = 1.2/K$ in the experiments. The term $1/K$ is a classical normalization factor and the constant 1.2 has been shown to yield reasonable sparseness (about 10 nonzero coefficients) in [11]. Since norm l_1 can better represent the distribution of the sparse coefficient and make the difference more distinguishable, γ_2 is then set to a relatively larger value (0.65) as the dominant weight, while γ_1 is 0.35. These two parameters are tested in practice as constants which are irrelevant to the noise level. C_1 and C_2 are defined as $(0.02C)^2$ and $(0.01C)^2$ where C is the dynamic range of the pixel values. In this paper, C is 255 for 8-bit grayscale images.

3. COMPARISONS AND EXPERIMENTAL RESULTS

3.1. Details of Implementation

To evaluate the performance of the proposed method, we tested the proposed method in two ways: one on a public dataset [15] with synthetic noise, and the other one on realistic videos captured under low light. The size of the two videos is 360×240. The sizes of the dictionary in the two-stage detection are 8×8 pixels with 256 atoms in the first stage and 3×3 pixels with 81 atoms in the second stage.

We add various levels of a mixture of noise to the public dataset [15]. The model of the mixture noise is defined as follows:

$$nI = \beta \bullet \left(\alpha P(I) + n \right) \tag{10}$$

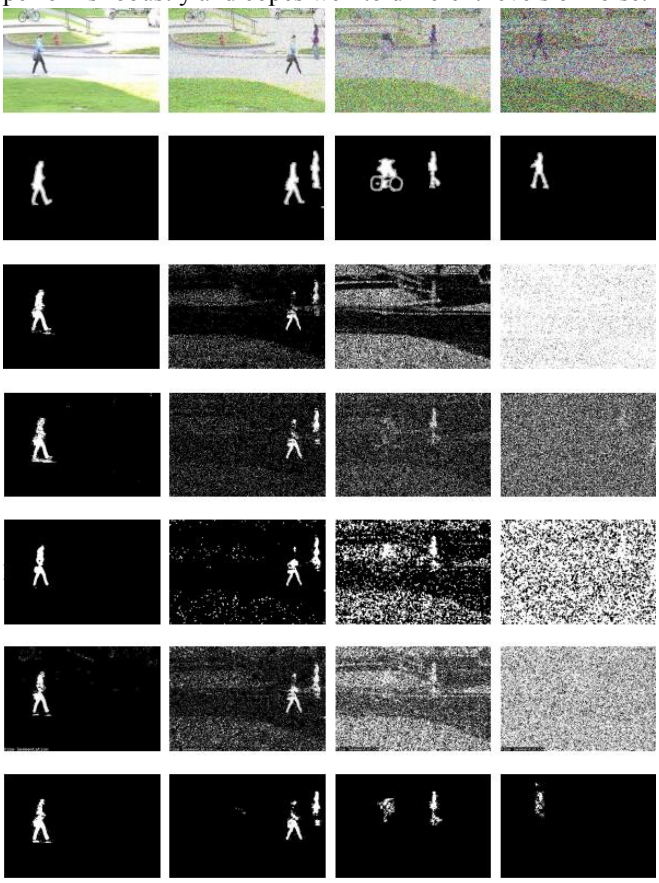
where I and nI are the original and noise image. α and β are the scale factor of Poisson noise and density parameter of salt & pepper noise. $P(\bullet)$ is the distribution of Poisson. n obeys the distribution of the Gaussian noise $N(\mu, \sigma^2)$. Equation (8) demonstrates that the mixture noise consists of Gaussian White, Poisson and salt & pepper noise.

3.2. Experimental results

In this section, we qualitatively compare the proposed method with classic background subtraction algorithms improved MoG [4] and KDE [5] and state-of-the-art algorithm SOBS [9] and ViBe [10]. For all these algorithms, we experiment with different adjustments of parameters until results seem optimal on the tested dataset.

First, we present qualitative comparisons over the public dataset [15] with synthetic noise. As shown in Fig. 3, this experiment extracts 4 frames (Frame No. 341, 419, 476, and 547) from dataset **Pedestrians** and adds three different levels of mixture noise (Noise level 1: $\sigma = 30$, $\alpha = 25$, $\beta = 0.01$; Noise level 2: $\sigma = 50$, $\alpha = 50$, $\beta = 0.02$; Noise level 3: $\sigma = 70$, $\alpha = 100$, $\beta = 0.03$). Under non-noise conditions (first column of Fig. 3), all these algorithms can achieve good detection result with little difference. When noise becomes high, the compared approaches are obviously infected by noise to different extent. The mixture noise at Level 1 affects KDE [5] the most and SOBS [11]

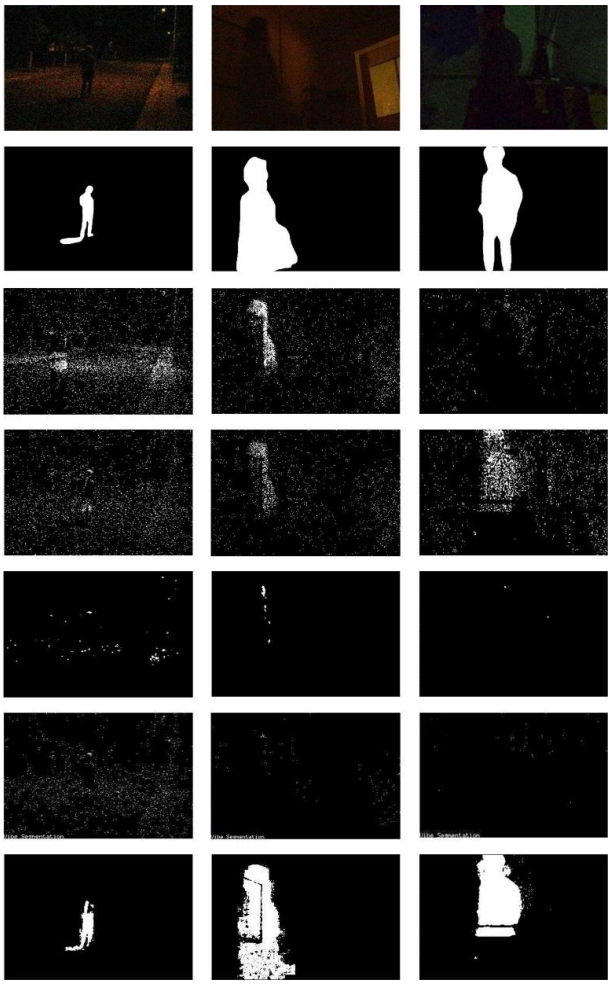
the least. At Level 3, the compared approaches lose efficacy almost completely. On the contrary, the proposed method performs robustly and copes well to different levels of noise.



(a)Original frame (b)Noise level 1 (c)Noise level 2 (d)Noise level 3

Fig. 3 Comparison of detection results on 4 frames (No. 341, 419, 476 and 547) from the dataset **Pedestrians** [15]. 1st row: images with synthetic noise [15]; 2nd row: ground truth [15]; 3nd to 7th row: detection results of improved MoG [4], KDE [5], SOBS [9], ViBe [10] and our method. Each column has different level of noise, from left to right (noise level 0: $\sigma = 0$, $\alpha = 0$, $\beta = 0$; noise level 1: $\sigma = 30$, $\alpha = 25$, $\beta = 0.01$; noise level 2: $\sigma = 50$, $\alpha = 50$, $\beta = 0.02$; noise level 1: $\sigma = 70$, $\alpha = 100$, $\beta = 0.03$)

Aside from the synthetic noise experiment, we implement the proposed method on different realistic low light videos with large noise as shown in Fig. 4. Here we use a SONY IMX 104 CMOS sensor. Under the low light condition, contrast enhancement is commonly used to enlarge the low grayscale value. However, it would increase the noise level which makes the detection even more difficult. Active infrared device is another alternative way to alleviate this problem, but it has limited distance and still captures corrupted videos only with less serious noise. Regardless of above methods, the proposed method are robust to handle harsh illumination environments.



(a)Dataset 1 (b) Dataset 2 (c) Dataset 3

Fig. 4 Comparison of detection results on real low light videos. 1st row: test frames extracted from low light videos. The illumination of the videos is about 1.5-2.0 **lx**, 0.7-1.0 **lx** and 0.3-0.5 **lx** from Dataset 1 to Dataset 3. 2nd-6th row: ground truth and detection results from Improved MoG [4], KDE [5], SOBS [9], ViBe [10] and our method.

4. CONCLUSIONS

Existing motion detection methods ignore the quality of image signals and are sensitive to noise. We propose a robust motion detection algorithm based on dictionary learning to handle noisy videos. When noise is too large for the assumptions of classic algorithms to hold, the mixture of Gaussian and non-parametric models becomes inappropriate, whereas the proposed method can still achieve satisfactory detection performance uninfluenced by statistical noise of different types and scales. Experimental results on synthetic and real noisy videos demonstrate the promising robustness of the proposed approach in comparisons with other competing methods. The real-time implementation of the proposed method can be considered the future direction of this work.

5. REFERENCES

[1] N. Friedman and S. Russell. "Image segmentation in video sequences: A probabilistic approach," *In Proc. of the 13th conference on Uncertainty in artificial intelligence*, pp. 175–181. Morgan Kaufmann Publishers Inc, 1997.

[2] C. Stauffer and W. E. L. Grimson. "Adaptive background mixture models for real-time tracking," *In Computer Vision and Pattern Recognition, IEEE Computer Society Conference*, 1999.

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

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

[5] A. Elgammal, D. Harwood, and L. Davis. "Non-parametric model for background subtraction," *In Computer Vision ECCV*, pp. 751–767. Springer Berlin Heidelberg, 2000.

[6] N. M. Oliver, B. Rosario, and A. P. Pentland. "A bayesian computer vision system for modeling human interactions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, 2000.

[7] V. Cevher, A. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa. "Compressive sensing for background subtraction," *In Computer Vision ECCV*, pp. 155–168. Springer Berlin Heidelberg, 2008.

[8] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh. "Background modeling and subtraction of dynamic scenes," *In Computer Vision, IEEE Int. Conf. on*, pages 1305–1312, 2003.

[9] L. Maddalena and A. Petrosino. "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Trans. on Image Processing*, vol. 17, no. 7, pp. 1168–1177, 2008.

[10] O. Barnich and M. V. Droogenbroeck. "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Trans. on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[11] J. Marial, F. Bach, J. Ponce, and G. Sapiro. "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research,* vol. 11, pp. 19–60, 2010.

[12] M. Aharon, M. Elad, and A. M. Bruckstein. "The k-svd: An algorithm for designing of overcomplete dictionarise for sparse representations," *IEEE Trans. on Signal Processing*, vol. 54, pp. 4311–4322, 2006.

[13] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[14] Z. Wang, A. C. Bovik, and H. R. Sheikh and E. P. Simoncelli. "Image quality assessment:From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[15] Dataset available: www.changedetection.net.