A Low Illumination Environment Motion Detection Method based on Dictionary Learning

Huaxin Xiao Yu Liu, Bin Wang, Shuren Tan and Maojun Zhang
College of Information System and Management, National University of Defense Technology, Changsha, P.R. China

Keywords: Low illumination, Motion detection, Dictionary learning, Sparse coding.

Abstract: This paper proposes a dictionary-based motion detection method on video images captured under low light with serious noise. The proposed approach trains a dictionary by background images without foreground. It then reconstructs the test image according to the theory of sparse coding, and introduces the Structural Similarity Index Measurement (SSIM) as the detection standard to identify the detection caused by the brightness and contrast ratio changes. Experimental results show that compared to the mixture of Gaussian model and ViBe method, the proposed method can reach a better result under extreme low illumination circumstance.

1 INTRODUCTION

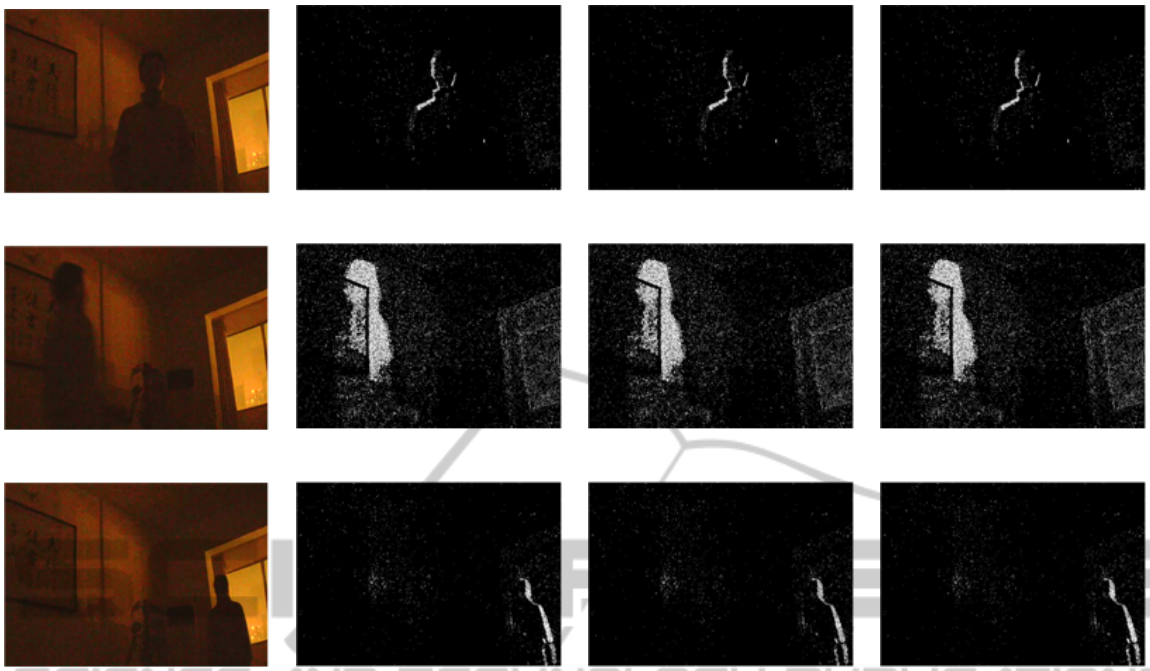
With the continuous improvement of equipment manufacturing and computer processing capability, the intelligent video surveillance technology has been widely applied to transportation, industry, defense and other fields. The intelligent processing in video surveillance such as tracking, classification, behavior understanding and so on, depends on the correct motion detection. Therefore, motion detection is a basic and crucial step with important research significance.

For a fixed scene camera, the commonly used detection methods are frame difference (Hui and Siu, 2007) and background subtraction (Piccardi, 2004). Frame difference method is fast, but for complex scenes the accuracy of detection is relatively low. Study and application on background subtraction are more widely. The main idea of this algorithm is to establish a background model of the monitored scene through a suitable method, then calculate the difference between the current frame image and the background model which segments the foreground area from the scene. Three Gaussian distributions corresponding to the road, shadow and vehicle (Friedman and Russell, 1997) were used to model the traffic surveillance system. Then, a mixture of multiple Gaussian distributions (Stauffer and Grimson, 1999) was employed to model the pixels in the scene which was proved to be a better solution to the modeling of complex background. Unlike the mixture of Gaussian model, Oliver et al. (Oliver et al., 2000) took into account the spatial correlation and captured the eigenbackgrounds by the eigenvalue decomposition. They adopted the Distance From Feature Space (DFFS) as a detection criterion. An incremental principal component analysis (Mittal et al., 2009) and robust principal component analysis (Wright et al., 2009) were respectively introduced which fully considered the structural information of the image. It can effectively deal with the brightness and other global changes. Recently, a universal background subtraction called ViBe (Barnich and Droogenbroeck, 2011) combined three innovative mechanisms to obtain a faster and better performance relatively.

The aforementioned methods are mainly for the complex and dynamic scene in the background, such as rain, waves and shaking trees, other than the low illumination environment. Large noise, low value and small differences in grey level are the typical characteristics of low light images. Excessive large noise and low grey value bring great influence on detection, which lead to the existing motion detection algorithms work improperly. In Fig. 1, we compare Fig. 1(b) the mixture of Gaussian model and Fig. 1(c) the ViBe method with Fig. 1(d) the proposed method. Compared with Fig. 1(b) and Fig. 1(c), the proposed method can effectively reduce the noise caused by low illumination and make the detection more robust.

In the perspective of the image blocks, the paper establishes the dictionary for each block, then reconstructs the test image according to the theory of sparse coding and finally treats the difference between the reconstruction image and the denoising background image and SSIM threshold as the detection criterions.

Xiao H., Liu Y., Wang B., Tan S. and Zhang M..
A Low Illumination Environment Motion Detection Method based on Dictionary Learning.
DOI: 10.5220/0004753701470154
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 147-154
ISBN: 978-989-758-018-5



(a) (b) (c) (d)

Figure 1: Motion detection methods comparison under low light. Fig. 1(a): The test images. The range of illumination in the test images is 0.1-0.5 **lx**. Fig. 1(b): The detection results of the mixture of Gaussian model (Stauffer and Grimson, 1999). Fig. 1(c): The detection results of the ViBe method (Barnich and Droogenbroeck, 2011). Fig. 1(d): The detection results of the proposed method.

This paper takes the mean of the sum of the collected background images as a denoising image to ensure the consistency of each experiment. As a result of the dictionary learning and sparse coding, the method can obviously remove most of the noise, and effectively identify detection caused by the brightness, contrast and other factors.

The rest of this paper is organized as follows. Section 2 describes the basic principle based on three assumptions. Section 3 presents the proposed method. Section 4 shows the experimental results under different illumination circumstances. Section 5 concludes and discusses future possible research direction.

2 BASIC PRINCIPLE

According to the approximate description of the proposed method on Section 1, the approach can be simply divided into three parts: basis vectors acquisition by dictionary learning, image reconstruction with sparse coding and foreground detection. The principles of the three parts are based on the below assumptions that make a good explanation of the proposed method in theoretical aspects.

An arbitrary signal $x \in \mathbb{R}^n$ can be represented sparsely and linearly by a small number of atoms in the dictionary $D \in \mathbb{R}^{n \times k}$:

$$\hat{x} = \arg\min_{\alpha} \|x - D\alpha\|_2^2 \qquad st. \|\alpha\|_0 \leq t \qquad (1)$$

where k is the number of atoms and $\|\alpha\|_0$ is the L_0 norm, counting the number of nonzero entries in the vector.

Many experiments indicate that such sparse decomposition is very effective in the application of signal processing (Chen et al., 1998). After the vectorization, an image can also be seen as a signal and decomposed sparsely, as described in the following assumption:

Assumption 1: any of an image can be sparsely and linearly represented by using some specific basis vectors in image space.

The information contained in an image can be represented by the particular structures that are also regarded as basis vectors in the image space. So, in Fig. 2(a), we show using the basis vectors to represent an arbitrary background of a scene.

Sparse decomposition always hopes that the reconstruction signal could be as close as possible to the original signal. The changes of the structures of the

background will bring new information which means there is a moving target. Then the original sparse decomposition will not exist, and we should reselect the bases of the image space. The process is shown in Fig. 2(b). Based on the above analysis, we propose the second hypothesis:

Assumption 2: the foreground will lead to changes in the structures of background, and then make the backgrounds bases in the image space transform.

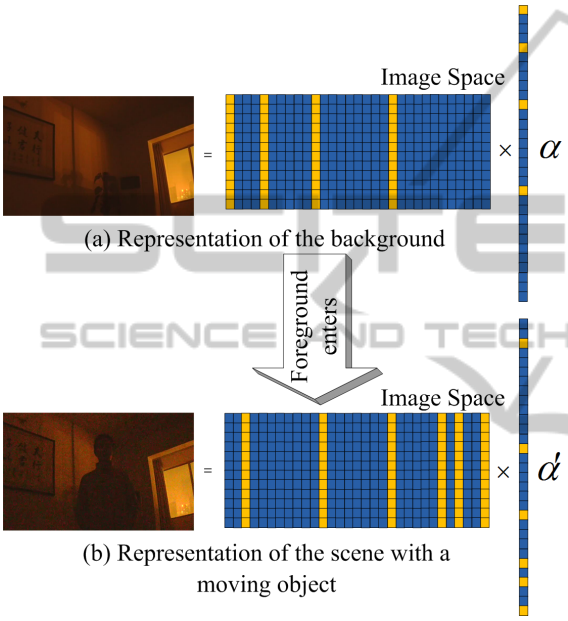
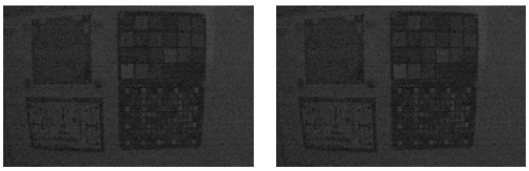


Figure 2: Sparse representation of the scene.

With the two assumptions, we can easily understand the proposed approach. When a background image with the foreground is reconstructed by the original bases, the part of the background which isnt affected by the foreground can be easily recovered. The other part, because of lack of the foreground bases, will be reconstructed with a deviation. Through the measure of the deviation, we can achieve the purpose of detection.

The two predominant sources of noise in digital image acquisition are the stochastic nature of the photon-counting and the intrinsic thermal and electronic fluctuations of the acquisition devices. With the decreasing of the light, the rapid boosting of the first one will lead to the surveillance video containing a large number of randomly distributed noise. When the noise flashing level is too large, it will make the existing detection methods not effective. Combination of the basis vectors idea, we consider the noise satisfies the following assumption:

Assumption 3: noise is randomly and discretely distributed in the basis vectors of image space.



(a) (b)

Figure 3: The result of denoising. Fig. 3(a): The original image with 0.05 **lx**. Fig. 3(b): The reconstructed image.

As described in Assumption 3, statistical noise is typically distributed through the larger space randomly and irregularly. So, when reconstructing an image though sparse coding, only several atoms in dictionary are selected to represent the original signal. Additionally, most of the noise can be effectively removed. These factors ensure our method can be well suited for handling low illumination environments. As shown in Fig. 3, the Fig. 3(a) contains large noise caused by the low illumination. After the dictionary learning and sparse reconstruction, the Fig. 3(b) eliminates the noise level, and well preserves the details.

3 THE PROPOSED METHOD

Those three assumptions described in Section 2 are the bases for the proposed approach. First, based on Assumption 1, we use the way of dictionary learning to obtain the basis vectors of the image space, and sparse coding to sparsely reconstruct the test image. Then according to Assumption 2, when the foreground object enters, it changes the structures of background and that part of recovered image. We can then combine the threshold region and SSIM values as the detection standards to determine the foreground area. In Fig. 4, we draw a brief flowchart about the process of dealing with each image block in three parts.

3.1 Dictionary Learning

Dictionary has been proved very effective for signal reconstruction and classification in audio and image processing domain (Mairal et al., 2010). Compared with the traditional methods such as wavelet and principal component analysis, dictionary learning does not emphasize the orthogonality of bases, which makes its representation of the signal have better adaptability and flexibility.

We extract the background frames without foreground from the surveillance video to form a training

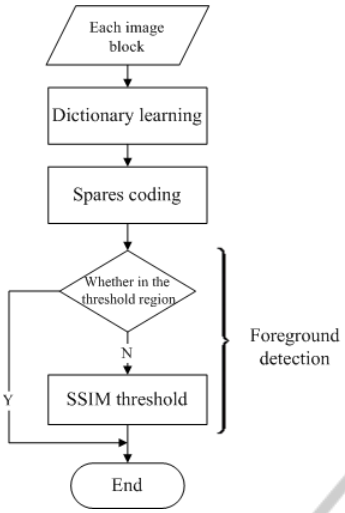


Figure 4: The brief flowchart of processing each image block.

set with N samples. As shown in Fig. 5, the collected images are divided into $m \times l$ blocks of size $\sqrt{n} \times \sqrt{n}$ pixels. The jth image block of the ith sample can be vectorized as a vector $\vec{x}_{ij} \in \mathbb{R}^n$. Then put the jth image block of each sample together and consist of a training set $X_j = \{\vec{x}_{ij} \mid i = 1, \cdots, N\}$ for the ith block. Its dictionary $D_j \in \mathbb{R}^{n \times k}$ satisfies the following formula:

$$D_j = \arg\min_{D_j} \sum_{i=1}^{N} \min_{\alpha_i} (\|\vec{x}_{ij} - D_j \alpha_i\|_2^2 + \lambda \|\alpha_i\|_1) \quad (2)$$

where α_i is the ith sparse coefficient and λ is a regularization parameter.

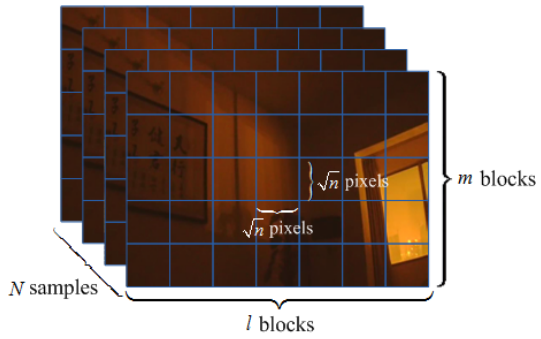


Figure 5: Creating the training set with N samples. Each image is divided into $m \times l$ blocks of size $\sqrt{n} \times \sqrt{n}$ pixels.

We use the Online Dictionary Learning algorithm (Mairal et al., 2010) to solve the formula (2). In each loop, the algorithm adopts stochastic gradient descent method to choose a vector $\vec{x}_{ij}$ which is regarded as x_t from X_j and t is the times of the repeat. Based on the previous $t-1$ loops, it applies sparse coding to get the tth decomposition coefficient α_t. The formula is as follows:

$$\alpha_t = \arg\min_{\alpha \in \mathbb{R}^k} \frac{1}{2} \|x_t - D_{t-1}\alpha\|_2^2 + \lambda \|\alpha\|_1 \quad (3)$$

Then update the dictionary $D_{t-1} = [d_1, \cdots, d_k$ column by column and get a new dictionary D_t. Update rules are as follows:

$$\begin{cases} u_j \leftarrow \frac{1}{A_{jj}}(b_j - D\alpha_j) + d_j \\ d_j \leftarrow \frac{1}{\max(\|u_j\|,1)} u_j \end{cases} \quad (4)$$

where $A = [a_1, \cdots, a_k] = \sum_{i=1}^{t} \alpha_i \alpha_i^T$ and $B = [b_1, \cdots, b_k] = \sum_{i=1}^{t} x_i \alpha_i^T$. The new dictionary meets the penalty function $\hat{f}_t(D) = \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \|x_i - D_{t-1}\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1$ minimum.

The algorithm is simple, fast and suitable for large scale image processing, and Mairal et al.(2010) have shown that the algorithm can converge to a fixed point. The test videos in this article have 1280×720 and 720×480 two specifications. High image quality of the monitored scene will lead to a result of great data. Other classical dictionary learning algorithms such as K-SVD (Aharon et al., 2006) take too much time to train the samples that can not meet the requirements of this paper.

We calculate the dictionary of each block in turn and then obtain the whole image dictionary D:

$$D = [D_1, D_2, \cdots, D_{m \times l}] \quad (5)$$

3.2 Sparse Coding

Sparse coding is a class of methods choosing good basis vectors automatically for unlabeled data. It discovers basis functions that capture higher level features in the data (Lee et al., 2006). In Section 3.1, we obtain the basis vectors by the dictionary learning. According to Assumption 1, we can use this set of basis vectors to reconstruct any of the test image and get the sparse coefficient on the basis vectors through sparse coding. For any of the test images Y, we use the same way of block and vectorization and get the vectors of blocks $\{\vec{y}_i \in \mathbb{R}^n \mid i = 1, \cdots, m \times l\}$. For any $\vec{y}_i$, its sparse coefficients on the dictionary should satisfy the following constraint:

$$\min_{\alpha \in \mathbb{R}^n} \|\vec{y}_i - D_i \alpha_i\|_2^2 \qquad st. \|\alpha_i\|_1 \leq t \quad (6)$$

or

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \|\vec{y}_i - D_i \alpha_i\|_2^2 + \lambda_1 \|\alpha_i\| + \frac{\lambda_2}{2} \|\alpha\|_2^2 \quad (7)$$

The constraint is a Lasso or Elastic-Net problem and we adopt the LARS algorithm (Efron et al., 2004) to solve it. For an arbitrary signal $\vec{y}_i$, the algorithm first computes the covariance matrix $D_i D_i^T$ and $D_i \vec{y}_i$, and decomposes them to get the sparse coefficients with a Cholesky-based algorithm. When the solution is sufficiently sparse and the scale of the problem is reasonable, LARS algorithm is very effective. Furthermore, the solution has the exact precision and does not rely on the correlation of atoms in the dictionary unless the solution is not unique.

Solve all image blocks in turn with LARS algorithm and we can obtain the sparse coefficients α of whole image:

$$\alpha = [\alpha_1^T, \alpha_2^T, \cdots, \alpha_{m\times l}^T]^T \tag{8}$$

Then, the reconstructed image $\hat{Y}$ is

$$\hat{Y} = D \times \alpha = \left\{ \vec{y}_i \in \mathbb{R}^n \mid i = 1, \cdots, m \times l \right\} \tag{9}$$

(a) (b)

Figure 6: Sparse reconstruction. Fig. 6(a): The grey scale images of testing images. Fig. 6(b):The grey scale images of reconstructed images.

Fig. 6 is a contrast with before and after sparse reconstruction. Fig. 6(a) and Fig. 6(b) are the grey scale images of the test images and reconstructed images respectively. We notice that Fig. 6(b) can well reconstruct the area of the background without foreground, while the area with moving target has a significant change. Furthermore, compared with Fig. 6(a), Fig. 6(b) has an obvious denoising effect and effectively reduces the noise impact on the detection

3.3 Foreground Detection

Referring to the idea of the background subtraction method, the paper calculates the difference between the background image $\bar{Y}$ and the reconstructed image $\hat{Y}$ by blocks. It then sums them to be the vector Δ:

$$\Delta = \left\{ \sum_{j=1}^{n} (\bar{Y}_i(j) - \hat{Y}_i(j)) \mid i = 1, \cdots m \times l \right\} \tag{10}$$

where $\bar{Y}_i(j)$ and $\hat{Y}_i(j)$ are respectively the jth pixel of ith block in $\bar{Y}$ and $\hat{Y}$.

Then we use the threshold region E to judge the vector Δ. Within the region, it means the structure of the block does not change, i.e. no foreground accesses. On contrary, there is an object enters the scene. The paper assumes the data in vector Δ approximatively follows the Gaussian distribution. Therefore, we set the upper and lower limit of the threshold region E with 3σ criterion:

$$\begin{cases} \max E = \mu + 3\sigma \\ \min E = \mu - 3\sigma \end{cases} \tag{11}$$

where μ and σ are the mean and variance of the differences between background images and reconstructed images in the training set.

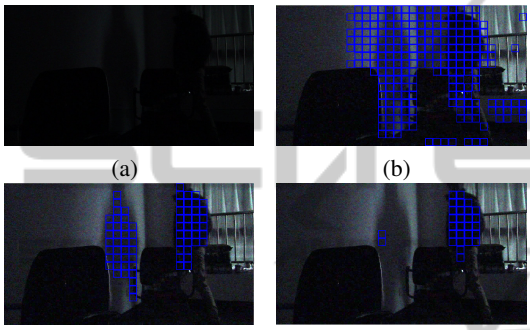
In low illumination environment, the sparse reconstruction cant distinguish the brightness, contrast and other information well. When only use the threshold region to determine whether the structure of the image block is changed, it is easy to cause wrong detection. To reduce the impact of non-structural information, this work introduces the SSIM as the detection criterion which defines the structure information independent of the brightness and contrast and reflects the properties of objects structure in the scene (Wang et al., 2004). The model of SSIM is defined as follow:

$$SSIM(X,Y) = [l(X,Y)]^{\alpha}[c(X,Y)]^{\beta}[s(X,Y)]^{\gamma} \tag{12}$$

where $l(X,Y)$, $c(X,Y)$ and $s(X,Y)$ are respectively the relative function of brightness, contrast and structure between block X and Y block. α, β and γ are the coefficients of weight.

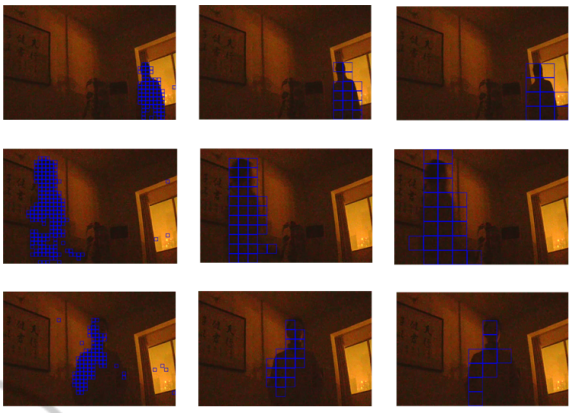
We judge these blocks not in the threshold region again with the SSIM threshold. Though Fig. 7(b), as a result of only using the threshold region, can effectively detect moving target, we can notice that there are many erroneous detection blocks with the structure that are hardly not changed. Calculate the SSIM

values of blocks in blue box and discover a large number of blocks SSIM values are above 0.8 that indicates many blocks without in the threshold region are similar to the ones in the background image. So we can take those blocks whose SSIM values are above a certain threshold into normal again. Fig. 7(c) and Fig. 7(d), as results with different SSIM thresholds, make the detection result more accurate. Especially in Fig. 7(d), it can filter out the shadow area because of covering. Since the original images illumination is too low, in order to facilitate observation, the paper increases the brightness of Fig. 7(b), Fig. 7(c) and Fig. 7(d).



(a) (b) (c) (d)

Figure 7: Present the detection results with blue boxes. Fig. 7(a): The original image with 0.01 **lx**. Fig. 7(b): Only using threshold region. Fig. 7(c): Using the threshold region and SSIM threshold and the threshold of SSIM is 0.9. Fig. 7(d): Using the threshold region and SSIM threshold and the threshold of SSIM is 0.8. In order to facilitate observation, the paper increases the brightness of Fig. 7(b), Fig. 7(c) and Fig. 7(d).



(a) (b) (c)

Figure 8: Different sizes of blocks comparison under 0.1-0.5 **lx**. Present the detection results with blue boxes. Fig. 8(a): 16×16. Fig. 8(b): 40×40. Fig. 8(c): 60×60.

(a) (b) (c)

Figure 9: The experimental results with 0.5-1.0 **lx**. Fig. 9(a): The grey scale images of reconstructed images. Fig. 9(b): The results indicated by the blue boxes. Fig. 9(c): The results of segmenting the foreground.

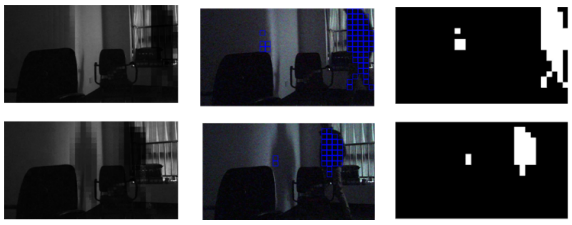
4 EXPERIMENTAL RESULT

In this paper, the image block is treated as processing unit and the size of blocks has a certain impact on the computing speed, detection results and recovered image effects.

Smaller blocks can ensure the precision of experimental results, as shown in Fig. 8(a). However, if the size of the block is over small, it is hardly to satisfy the accuracy of detection, as shown in the second row of Fig. 8. Larger blocks can guarantee the accuracy and have a better denoising effect with higher computing cost. Precision and accuracy are a pair of tradeoff parameters and it is difficult to simultaneously ensure both at a high level. After several tests, we respectively select 16×16 and 40×40 as the block size for resolution of 720×480 in Fig. 8 and 1280×720 in Fig. 9 and Fig. 10.

In order to fully test the proposed method, we use surveillance cameras to capture a number of videos at different illumination and background environments. In Fig. 8, 9 and 10, the range of illumination is respectively 0.1-0.5 **lx**, 0.5-1.0 **lx** and 0.01-0.05 **lx**. The (a) of Fig. 9 and 10 are the recovered grey scale images by using dictionary learning and sparse coding. Contrast with the background image, you can find the area with foreground entering significantly changed. (b) are the detection results through the threshold region and SSIM filtering out. Just as shown in Fig. 7, in order to facilitate observation, the paper increases the brightness of Fig. 10 (a) and (b). (c) is the result of dividing the foreground.

For extreme low illumination (under 0.1 **lx**) and grey level (less than 15) in Fig. 10, the proposed method can effectively detect the moving target. For



(a) (b) (c)

Figure 10: The experimental results with 0.01-0.05 **lx**. Fig. 10(a): The grey scale images of reconstructed images. Fig. 10(b): The results indicated by the blue boxes. Fig. 10(c): The results of segmenting the foreground. In order to facilitate observation, the paper increases the brightness of Fig. 10(a) and Fig. 10(b).

small objects suddenly appearing in the background, the present method can also effectively detect the target, as shown in Fig. 9. The small objects in the Fig. 9 are a foam cross and a paper box and take about 20 frames to across the whole scene. However, since the dictionary is learned from the background images without foreground, the recovered blocks are close to the background when the colours of foreground and background are similar. This circumstance increases the difficulty in detecting, such as the human legs in the second row of Fig. 10.

Table 1: Detection results under different illumination of environment and sizes of moving target. The size of the block is 40×40 pixels.

	1 **lx**	0.5 **lx**	0.1 **lx**	0.01 **lx**
5 blocks	100%	100%	40%	20%
10 blocks	100%	90%	70%	60%
30 blocks	89%	85%	83%	70%
50 blocks	86%	90%	80%	68%
100 blocks	85%	88%	79%	76%

In Table 1, we simply census the detection results under different illumination of environment and sizes of moving target. The left-most column of the Table 1 presents the number of blocks of the moving object occupying in the image and the size of the blocks is 40×40 pixels. These values are approximation. The top row in Table 1 shows the different values of illumination. The percent describes the proportion of the object that the proposed method can detect. We can find that when the illumination is above 0.5 **lx**, the proposed method can detect near 90% of the blocks of the moving target holding in the image. With the illumination decreasing to 0.01 **lx**, it can still identify about 70% of the blocks while it is difficult for human vision to distinguish most of them, such as in Fig. 7. The Table 1 sufficiently reflects the robust of the proposed method in extreme low illumination environment.

5 DISCUSSION

Most of the existing motion detection algorithms do not adequately take into account the extreme low illumination situation. This paper proposes a motion detection algorithm based on dictionary learning on video images captured under low light. The experiments show that compared to the mixture of Gaussian model and the ViBe method, the proposed method achieves a better detection results even in the case that human eyes are difficult to distinguish. When a portion of the moving object is close to the background, it is difficult to detect this region which is the inadequacy of this paper. In addition, the paper also carries out small objects detection under low light experiment. Smaller and faster motion detection in low illumination can be considered as the future direction of this work.

ACKNOWLEDGEMENTS

This research was partially supported by National Science Foundation of China (NSFC) under project No.61175006 and No.61175015.

REFERENCES

Aharon, M., Elad, M., and Bruckstein, A. (2006). K-svd: an algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322.

Barnich, O. and Droogenbroeck, M. V. (2011). Vibe: A universal background subtraction algorithm for video sequences. *Image Processing, IEEE Transactions on*, 20(6):1709–1724.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *scientific computing*, 20(1):33–61.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Friedman, N. and Russell, S. (1997). Image segmentation in video sequences: a probabilistic approach. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages :1232–1245.

Hui, K. C. and Siu, W. C. (2007). Extended analysis of motion-compensated frame difference for block-based motion prediction error. *Image Processing, IEEE Transactions on*, 16(5):1232–1245.

Lee, H., Battle, A., Raina, R., and Ng, A. (2006). Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages :801–808.

Mairal, J., Bach, F., and J. Ponce, G. S. (2010). Online learning for matrix factorization and sparse coding. *Machine Learning Research*, 11:19–60.

Mittal, A., Monnet, A., and Paragios, N. (2009). Scene modeling and change detection in dynamic scenes: a subspace approach. *Computer Vision and Image Understanding*, 113(1):63–79.

Oliver, N. M., Rosario, B., and Pentland, A. P. (2000). A bayesian computer vision system for modeling human interactions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):831–843.

Piccardi, M. (2004). Background subtraction techniques: a review. In *Systems, Man and Cybernetics, IEEE international conference on*, pages :3099–3104.

Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, IEEE conference on*, pages :246–252.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612.

Wright, J., Ganesh, A., and S. Rao, Y. M. (2009). Robust principal component analysis? In *Proceedings of the Conference on Neural Information Processing Systems*.