IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 October 2023, accepted 27 October 2023, date of publication 2 November 2023, date of current version 8 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329816

RESEARCH ARTICLE

A Novel Background Modeling Based on Keyframe and Particle Shape Property for Surveillance Video

YONG FAN[1,2], XIU HE[1], YIYI LIN[1], AND ZHANCHUAN CAI[1], (Senior Member, IEEE)

[1]School of Computer Science and Engineering, Macau University of Science and Technology, Macau, China
[2]Computer College, Guangdong University of Science and Technology, Dongguan 523668, China

Corresponding author: Zhanchuan Cai (zccai@must.edu.mo)

This work was supported in part by the Science and Technology Development Fund of Macau under Grant 0052/2020/AFJ and Grant 0059/2020/A2, in part by Zhuhai Industry-University-Research Collaboration Program under Grant ZH22017002210011PWC, in part by the Dongguan Science and Technology of Social Development Program under Grant 20211800904192, and in part by the Research Project of Guangdong University of Science and Technology under Grant GKY-2020KYYBK-23.

ABSTRACT With the development of industrial informatization, video processing technology is receiving more and more attention. Extracting background is a prerequisite for many video processing techniques, so video background modeling technology is becoming highly sought-after. Currently, there are a variety of approaches to estimating background; however, many of these methods have the fault of not being able to accurately distinguish between foreground and background, especially when objects move slowly or remain still for a period of time. In this paper, a novel background modeling scheme is proposed for surveillance video, based on keyframe and particle shape properties. The model consists of three parts: the first part is to reduce running time and eliminate the ghost phenomenon caused by adjacent redundant frames by extracting keyframe and dividing the extracted frames into several groups; the second part includes three steps, computing binarized difference, characterizing the binarized difference and screening the difference where a quadruple, composed of particle shape properties, is designed to quantitatively describe binarized differences; the third part involves generating the temporary background and updating the temporary background according to the similarity of data obtained at the newly proposed locations. Experiment results on SBMnet and SBI datasets and comparisons with some emerging algorithms show that the performance of the proposed model is superior or comparable to the other state-of-the-art methods particularly when dealing with stationary objects. Furthermore, the proposed method ranks as the second for intermittent category video, compared to the other 31 state-of-art methods. Moreover, the speed of the proposed method, 5.38 **F**rames **P**er **S**econd (FPS) for SBMnet dataset and 10.94 FPS for SBI dataset, is faster than most public methods.

INDEX TERMS Video processing, surveillance video, background modeling, difference, quadruple, keyframe, particle shape, stationary object.

I. INTRODUCTION

With the informatization of industry and the increase in the number of surveillance cameras, there is a strong demand for the automatic processing of captured videos. In this field, background estimation is a basic, low-level task applied as a pre-processing step for object tracking, video compression, inpainting, privacy protection, computational photography and so on [1], [2], [3], [4], [5], [6], [7], [8]. Aiming at the problems of occlusion, drift, and background change in visual image tracking, Ren et al. [1] presented a background learning correlation filtering algorithm based on multi-feature fusion. In the framework of correlation filtering, multi-feature fusion, multi-template update, and background learning regularization are used to improve the performance of the filter in the problem of template contamination and

The associate editor coordinating the review of this manuscript and approving it for publication was Yun Zhang.

object occlusion. In [2], Lu and Huang proposed a novel sparse-representation based hyperspectral anomaly detection method via adaptive background sub-dictionaries. In the paper, a background estimation strategy is proposed to provide representative background information. Based on the estimated background, a global dictionary is constructed by utilizing K-means clustering algorithm. In practical missions, although general sketch of the background is stable, some details change constantly. Aimed at this, Pei et al. [3] proposed to represent the general background by a linear combination of some atoms and record the detailed background by spatiotemporal clustered patches.

Kim et al. [4] proposed an integrated network that produces two kinds of outputs a background model image and a foreground object map to adapt to the new environment by retraining using a background model image. In the realm of industrial surveillance environments, Lyu et al. [5] proposed a visual early leakage detection system that employs an established background model to extract dynamic potential leakage foreground. Similarly, Ma et al. [6] presented a novel approach to expedite the search process of surveillance video coding through the utilization of a background model. Their method involves an initial step of background modeling, followed by the implementation of ''coding units classification'' based on the established background. Wang et al. [7] emphasized the importance of surveillance video coding in improving compression efficiency in intelligent video surveillance systems and applications. They proposed a background modeling and referencing scheme for moving cameras-captured surveillance video coding in **H**igh-**E**fficiency **V**ideo **C**oding (HEVC). The scheme includes a low-complexity motion background modeling algorithm for surveillance video coding and the use of motion background coding tree units to update the previous coding tree unit in the global compensation location of the background reference picture. Experimental results demonstrated significant bit savings of up to 26.6% and, on average, 6.7% with similar subjective quality and negligible encoding complexity compared to HEVC reference software HM12.0. Tezcan et al. [8] introduced a new, supervised, background subtraction algorithm for unseen videos based on a fully-convolutional neural network. They argued that the success of deep learning in computer vision did not bypass the background subtraction algorithm, which is founded on the results of background modeling. The diverse range of applications of background modeling technology in the domain of video processing has led to significant research efforts by scholars in this area.

Background modeling, also referred to as background estimation, reconstruction, extraction, bootstrapping, or generation, is a process aimed at effectively and precisely retrieving a background devoid of any foreground objects (moving or stationary objects) from a sequence of frames [9]. Generally speaking, there is no formal definition of background or foreground. Certain assumptions are typically made in order to make the task feasible. In [10], the background is defined as being globally stationary, and in each pixel, the background is visible for at least a short interval of the sequence. Additionally, only foreground objects may be in motion during this period. However, in real-world environments, the definition of foreground and background may be significantly impacted by various issues, such as scenarios in which foreground objects remain static for extended periods of time before beginning to move (e.g., parked cars or stopped pedestrians that suddenly start to move). This poses the biggest challenge on the account that any change detection solution that does not focus explicitly on static object detection and segmentation [11]. During the last two decades, many methods had been proposed in order to address these tasks in the context of moving objects detection [12]. Liu et al. [13] proposed a framework for scene background modelling based on temporal median filter with Gaussian filtering. However, the proposed method has the drawback of mistakenly distinguishing the foreground from the background when objects move slowly or stay still for some time. Mseddi et al. [14] proposed NExBI method based on online block-level processing to initialize the background. The method is used to solve the task of clutter scenes, and results for other categories show some somewhat unsatisfactory performance. In 2021, Li et al. [15] proposed a new non-convex sparsity model based on background subtraction. The method only provides results conducted on the clutter category of SBMnet. The results for other types of videos are unknown. In 2022, Sauvalle and Fortelle [16] proposed a new method for fixed background reconstruction using stochastic gradient descent. However, this method still does not solve the problem that objects are mistakenly classified as backgrounds. In recent years, with the profound study of machine learning, many background modeling methods based on machine learning have Aemerged. Halfaoui et al. [17] proposed a solution to estimate the initial background based on a **C**onvolutional **N**eural **N**etwork. Sultana et al. [18] proposed an end-to-end framework based on a **G**enerative **A**dversarial **N**etwork to handle the problem of dynamic background modeling. Zhao et al. [19] proposed a universal background subtraction framework based on the **A**rithmetic **D**istribution **N**eural **N**etwork for learning distributions during background subtraction. It is widely acknowledged that various algorithms have distinct strengths and weaknesses, and their appropriateness depends on the nature of the problem being addressed. Therefore, careful selection and optimization of algorithms are crucial for achieving desired outcomes. As this is not a comprehensive survey paper, it is essential to note that only conventional methods employed for video background reconstruction will be addressed in this study.

Even though many approaches have been proposed to reconstruct backgrounds; however, the performance of these methods still encounters challenges when dealing with videos containing slow-moving or stationary foreground objects. In real world scenarios, the technologies that are

able to extract background without foreground are strongly required in video surveillance systems. To address the issue, a novel background model, called **B**ackground **M**odeling based on **K**eyframe and **P**article shape properties (BM-KP), to automatically reconstruct the background without prior knowledge is proposed. The model consists of three parts. The first part is to extract keyframes and divide them into several groups, in order to reduce the running time and get rid of the redundant frames. The second part includes three steps: computing binarized difference, characterizing the binarized difference and screening the difference. In this part, the quadruples are devised to offer a quantitative depiction of binarized differences and eliminate unsuitable binarized differences. The third part consists of generating the temporary background and updating the background according to the similarity of data obtained at the newly proposed locations. As reported in [20], our proposed method can be classified as methods based on iterative model completion from the view of methodology, hybrid methods(operating at both the pixel and region levels), and offline methods based on the whole sequence. Qualitative and quantitative results demonstrate that the performance of the proposed model is superior or comparable to the other state-of-the-art methods particularly when dealing with stationary objects, while having faster speed.

The main contributions of this paper are as follows. 1) **The pre-processing process of keyframe extraction is utilized to address the ghost phenomenon and reduce running time.** The pre-processing process eliminates adjacent redundant frames and selects representative frames to capture the essential content of the video. Compared to other algorithms, this process effectively suppresses the formation of ghosts caused by overlapping targets in adjacent frames and has a faster speed (More details can be found in Fig. 5 and Table 9). **2) The scheme of quantitative description for binarized differences is proposed.** It involves treating the difference between two frames as particles and the quadruple consisting of particle shape properties, which provides a way of quantitatively describing the frame difference. Based on this scheme, it is easier to select the desired frame differences, which can be beneficial for many applications such as video surveillance (more details can be found in Section III-C). **3) The new locations used for similarity comparison are proposed.** The key task of background updating is to decide whether new pixels or blocks of candidate frames belong to the background according to some similarity criteria. Previous algorithms typically performed similarity comparisons between adjacent blocks. However, the single location of comparison presents challenges in real world environments. To address this issue, it is suggested that the external area adjacent to the boundary, the boundary itself, the interior area adjacent to the boundary, the center of the object, and the interior area of the mask be used for similarity comparison. This strategy is more inclusive and robust compared to previous algorithms when dealing with complex scenarios, particularly those involving stationary objects (More details can be found in Section III-E). **4) The process consisting of several filtering to select candidate frames is proposed to avoid the use of a unique function.** In previous algorithms, the candidate having the most smoothness measured by one single function is labelled as background. Unfortunately, this is not always true, especially when the neighboring blocks are occluded by stationary objects and the border of blocks happens to have the same color as stationary objects. Therefore, in order to address the issue, it is recommended to employ a process of filtering frames in a stepwise manner, starting from coarser to finer levels, utilizing the data obtained from the proposed locations. Experiments show that by implementing this process, the proposed model avoids selecting the wrong candidate frames when there are stationary objects in frames (more details can be found in Section III-E and Fig. 5).

The rest of this paper is structured as follows. Section II discusses existing approaches that are relevant to our work. Section III provides a detailed description of the key steps of the proposed model. Section IV presents and discusses the results obtained from the implementation of our proposed model. Finally, Section V presents the conclusion and future work.

II. RELATED WORKS

The topic of background modeling is vast and has been widely researched by many scholars [9], [10], [13], [14], [15], [16], [17], [18], [19], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33]. Different terms have been used to refer to background modeling, including background estimation, bootstrapping, background initialization, background generation and background reconstruction etc [20]. However, there is not an unique categorization of background modeling methods. In [34], these methods are classified according to different aspects. In this section, region-based methods, pixel-based methods, hybrid methods, iterative model completion-based methods, and online-offline methods that are relevant to our works are discussed (For more details about background modeling methods, it is advisable to read [12], [20], [34]).

Region-based methods exploit spatial relations by partitioning the image into blocks and constructing a background model for each image block. However, since motion detection is performed at the patch level, misclassified blocks can propagate errors to the pixel level, resulting in low accuracy. Pixel-level methods process each pixel independently, but they are sensitive to ‘‘ghosting’’ artifacts. Hybrid methods operate at both the pixel and region levels, thus providing a balance between efficiency and accuracy. Methods based on iterative model completion construct the background in an iterative manner, by first identifying areas where no activity has been detected, which are then used as background initialization. From there, the background model is iteratively completed based on some criteria. Online methods process frames one by one, without going back in time, and they have the disadvantage of always incorporating changes that occur

in the background. Offline methods compute the background by considering the entire video frames as a whole and often have memory issues when the video is long. In summary, despite many methods have been proposed to estimate background, the main limitation that existing schemes do not successfully deal with stationary objects remains unsolved.

In the current study, a novel model, which draws inspiration from **keyframe, frame difference and particle shape**, is proposed to address the aforementioned limitations. 1) Keyframes, which reflect the main content of a video shot, can be used more effectively than the original video streams as indexes for video streams [35]. In our work, an improved keyframe extraction method is proposed to reduce running time and mitigate the ghost phenomenon caused by adjacent redundant frames (More details can be found in Section III-B). 2) The basic principle of the difference method is to identify objects by adjacent frame subtracting. The method involves initially subtracting adjacent frames, followed by comparing the subtracted outcomes with a threshold. Through the application of binarized frame difference, it becomes possible to identify the regions in the two frames where the difference exceeds a specified threshold, thereby classifying them as foreground targets. Conversely, the areas falling below the threshold are designated as background. Subsequently, in subsequent stages, additional operations such as morphology can be readily conducted on the binarized difference. Notably, the utilization of binarized difference enables a faster distinction between the foreground and background compared to alternative methods, such as the optical flow approach. Assuming that the input frame sequence is noted as $F = \{f_i | i = 1, 2, 3, \ldots, n\}$, where i is the sequence number, and n is the length of the sequence. The subtracting operation can be described as Eq.(1):

$$d_i(x, y) = |f_{i+1}(x, y) - f_i(x, y)| \tag{1}$$

The binarized diffenece, which is obtained by comparing the frame difference with the threshold, can be represented as Eq.(2)

$$b_i(x, y) = \begin{cases} 1, & if \quad d_i(x, y) \geqslant th \\ 0, & otherwise \end{cases} \tag{2}$$

where $d_i(x, y)$ is the difference between the $(i + 1)th$ frame and the ith frame at location (x, y) $(1 \leqslant x \leqslant w, 1 \leqslant y \leqslant h)$, and w, h represents the width and height of the frame, respectively. $b_i(x, y)$ is the binarized value at (x, y) by comparing $d_i(x, y)$ with the threshold th, which is automatically calculated by Ostu method [36]. As of now, a variety of difference methods are proposed to detect the moving object. Despite their high computation, they fail in dealing with frames where foreground objects move slowly or become stationary. 3) Shape is a fundamental property of all objects. In [37], shape properties such as sphericity, roundness, irregularity and roughness are defined to describe sedimentary particles. **In this study, a novel scheme is proposed to treat the differing portion of the binarized difference as particles.** This scheme utilizes the particle attributes to characterize the differing portion of difference. Consequently, a quantitative representation of binary frame differences can be attained. As a result, the suitable frame can be easily chosen to establish the initial background. (More details can be found in Section III-C).

III. THE PROPOSED METHOD

In this section, a detailed explanation of the proposed **BM-KP** method is presented. An overview of the proposed method is provided in Section III-A, followed by a thorough description of each component from Section III-B to Section III-E.

A. AN OVERVIEW

The proposed model consists of three parts. The first part involves the extraction of keyframes and the division of these extracted frames into several groups. This step is intended to reduce the running time by extracting representative frames. The second part entails three steps: computing binarized difference, characterizing the binarized difference and screening the difference. This part is designed to remove binarized differences in which objects are mixed up with background. The third part involves the generation of a temporary background and its updating. In this part, suitable frames are selected to update the temporary background-in other words, the unknown parts under masks of the temporary background are estimated based on the similarity of input frames and the temporary background at the new proposed locations. The flowchart of the proposed model is presented in Fig.1, and the steps are described in detail below. Furthermore, to facilitate comprehension, Table 1, which outlines the main symbols and corresponding explanations utilized throughout the paper, is provided.

B. EXTRACTING AND GROUPING KEYFRAMES

As previously mentioned, the initial step of the proposed model is to eliminate redundant frames through iterative extraction of keyframes. One example is illustrated in "Extract Keyframe" part of Fig.1. Furthermore, Algorithm 1 outlines the pseudo code utilized for extracting keyframes. Assuming that the similarity between adjacent frames is represented as

$$s_i = \sqrt{\sum_{x=1}^{w}\sum_{y=1}^{h}(d_i(x, y) - \overline{d_i})^2 \Big/ (wh - 1)} \tag{3}$$

where $\overline{d_i}$ is the average of all $d_i(x, y)(1 \leqslant x \leqslant w, 1 \leqslant y \leqslant h)$, and the sequence of s_i is represented as $S = \{s_i | i = 1, 2, 3, \ldots n - 1\}$. Essentially, s_i denotes the standard deviation derived from the differences observed across all pixel positions in two frames. Consequently, utilizing s_i as

TABLE 1. The main symbols and corresponding explanations.

Symbol	Explanation
F, n	The sequence of input frames and its size.
$d_i(x, y)$	The difference between the $(i+1)th$ frame and the *ith* frame at location (x, y).
$b_i(x, y)$	The binarized value obtained by conducting a comparison between $d_i(x, y)$ and the threshold *th*.
w, h	The width and height of frame.
s_i	The similarity between frames.
S	The set of s_i.
K, η	The sequence of extracted keyframe and its size.
λ, K_i	The interval of dividing the keyframe sequence K into subsequences, and the *ith* subsequence.
$\Delta^i_{j,k}$	The binarized difference obtained by the *jth* keyframe and *kth* keyframe in K_i.
V^i_j	A set formed by the frame difference between the *jth* frame in set K_i and the remaining frames in the set.
$\rho^i_{j,k}$	The quadruple used to quantitatively describe $\Delta^i_{j,k}$.
$\alpha^i_j(k)$	To indicate whether the binarized differences $\Delta^i_{j,k}$ is retained..
B^i_j	The temporary background reconstructed by the *jth* keyframe with the other keyframes in subsequence K_i .
B^i	The temporary background reconstructed by all keyframes in K_i.
B	The temporary background reconstructed by all subsequence K_i.
M^i_j, M^i, M	The mask matrix for B^i_j, B^i and B respectively.
M_i	The *ith* mask in mask matrix M.
A_i, φ	The size of *ith* mask and the block size.
L^b_i, L_i, L^e_i, L^i_i	The boundary locations of M_i, the interior locations of M_i , the external locations adjacent to the boundary, and the interior locations adjacent to the boundary.
$\chi_{i,\iota}$	The six data sets constructed with local information to assess the local similarity, where, $\iota = 1, 2, 3, 4, 5, 6$
$\gamma^k, \gamma^k(\chi_{i,\iota})$	The threshold used to select candidate frames in the *kth* round selection, and the threshold for set $\chi_{i,\iota}$ in the *kth* round selection.
ϕ^k_ι	A constant related to the category of $\chi_{i,\iota}$ and the round of selection. For the first selection, the values of ϕ^1_1, ϕ^1_2, ϕ^1_3, ϕ^1_4, ϕ^1_5 and ϕ^1_6 are set to -0.75, 1, 0, 0.5, 1, and 0.25, respectively. For the second selection, they are set to 1, 0.25,0.75, 1, 1, and 1 respectively.
θ_j	It is used to represent whether the *jth* frame is reserved to reconstruct the background in the updating temporary background stage.

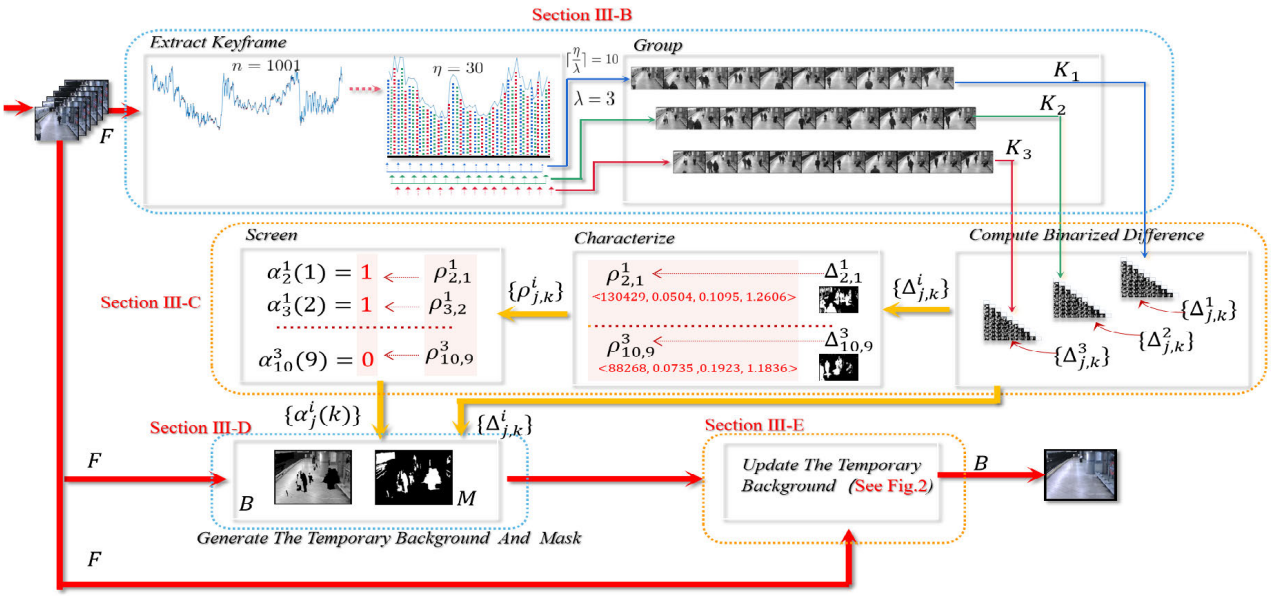


FIGURE 1. The flowchart of the proposed model. Initially, keyframes are extracted and grouped. Then, binarized differences between any two keyframes in each group are computed and digitally described using a 4-tuple. These binarized differences are screened according to the Pauta Criterion. Subsequently, a temporary background, potentially containing masks for stationary objects, is generated using the selected binarized difference and input frames. Finally, backgrounds under the masks are updated according to the process proposed in this paper(Details in Fig.2).

a representation of similarity between the frames is justified. Furthermore, a smaller s_i value indicates a higher degree of similarity between the two frames. The functions employed for obtaining frames characterized by local maximum values of s_i, frames characterized by local minimum values of s_i, and frames with s_i surpassing the mean value of all s_i, are

TABLE 2. A comparative analysis of the proposed model and BI-GAN [18] algorithm in Age, pEPs, pCEPS, and MSSSIM scores on six videos of SBI dataset. (The red indicators are the ones that the BI-GAN [18] algorithm wins, while the rest of the algorithms are the ones that the suggested model wins.)

Video	AGE↓		pEPs↓		pCEPS↓		MSSSIM↑		PSNR↑	
	BM-KP	BI-GAN [18]	**BM-KP**	BI-GAN [18]	**BM-KP**	BI-GAN [18]	**BM-KP**	BI-GAN [18]	**BM-KP**	BI-GAN [18]
Candela_m1.10	2.8876	4.9075	0.2308	0.9033	0.0000	0.0488	0.9956	0.9800	35.4889	31.3445
CAVIAR1	2.4832	8.5872	0.1689	7.0801	0.0793	0.8057	0.9948	0.9529	37.4907	26.2490
CAVIAR2	0.8517	12.7988	0.0132	13.6963	0.0031	2.0020	0.9991	0.9809	45.8774	24.8151
CaVignal	1.8810	4.5303	0.0184	2.1484	0.0000	0.0977	0.9957	0.9877	39.0900	30.8977
HallAndMonitor	2.5275	3.9255	0.2344	0.1709	0.0781	0.0000	0.9853	0.9899	34.0840	34.1297
HumanBody2	4.6120	7.1523	0.5143	4.1748	0.0000	0.2930	0.9950	0.9744	32.4432	28.7653

Algorithm 1 Extraction of Keyframe

Input: Sequence $F = \{f_i | i = 1, 2, 3, \ldots, n\}$
Output: Sequence of extracted keyframes K, and its size η

$K \leftarrow F$, $\eta \leftarrow$ n;
$\zeta = 20$; ▷ Control the scale of extracted keyframes
if $\eta \geq \zeta$ **then**
  $K = ExtractKeyFrame(K, 1)$;
  $K = ExtractKeyFrame(K, 2)$;
  $K = ExtractKeyFrame(K, 3)$;
  $\eta \leftarrow$ the size of K;
  while $\eta > \zeta$ **do**
    $K = ExtractKeyFrame(K, 1)$;
    $K = ExtractKeyFrame(K, 2)$;
    $\eta \leftarrow$ the size of K;
  end while
end if
return K and η.

function ExtractKeyFrame(K,c)
  $\eta \leftarrow$ the size of sequence K;
  for i=1 to η-1 **do**
    Obtain $d_i(x, y)$ of the sq by Eq.(1);
    Obtain s_i based on $d_i(x, y)$;
  end for
  $S \leftarrow \{s_i\}$;
  if $c == 1$ **then**
    $K = g_1(S)$;
  else if $c == 2$ **then**
    $K = g_2(S)$;
  else if $c == 3$ **then**
    $K = g_3(S)$;
  end if
return K
end function

respectively defined as follows:

$$\begin{aligned} g_1(S) &= \{f_i | s_i \geqslant s_{i-1}, s_i \geqslant s_{i+1}\} \\ g_2(S) &= \{f_i | s_i \leq s_{i-1}, s_i \leq s_{i+1}\} \quad for \quad \forall i \in [1, n] \\ g_3(S) &= \{f_i | s_i \geqslant \bar{s}\} \end{aligned} \tag{4}$$

where $\bar{s}$ is the mean of all s_i, and i is the sequence number of frame. $g_1()$, $g_2()$, and $g_3()$ are three functions devised with the objective of retrieving frames that satisfy the predetermined conditions. Assuming the last sequence of extracted keyframes is represented as K, the number of extracted keyframes, denoted as η. Algorithm 1 utilizes the ExtractKeyFrame(K,c) subfunction to extract frames, with ζ serving as the parameter to regulate the scale of the extracted frames. These chosen frames are subsequently arranged into a sequence of keyframes denoted as K. This iterative process persists until the length of the keyframe η becomes smaller than the specified threshold, ζ.

Afterwards, the keyframe sequence, K, is segmented into subsequences at regular intervals of λ. This segmentation process further diminishes the similarity between frames within each sub sequence. As a result, these subsequences can be represented as:

$$K_i = \{f_{i'} | f_{i'} \in K\} \tag{5}$$

where $1 \leqslant i \leqslant \lambda$, i' is the sequence number of frame in the last extracted keyframe sequence, K, and it satisfies $i' \in \{i + \lambda(m-1)\}_{1 \leqslant m \leqslant \lceil \frac{\eta}{\lambda} \rceil}$. m is the sequence number of extracted keyframe in subgroup K_i.

C. CHARACTERING AND SCREENING BINARIZED DIFFERENCE

In this step, in order to reduce the possibility of generating incomplete edges and ghost, a strategy was designed for converting frame difference into quadruple. Next, the inappropriate binarized differences are filtered out in accordance with the devised strategy. Assuming that the binarized difference between the jth and kth keyframe in K_i is noted as $\Delta^i_{j,k} (k \neq j, 1 \leqslant j, k \leqslant |K_i|)$, where $|K_i|$ represents the size of subsequence K_i, the set of the binarized difference obtained by the jth keyframe with the others in K_i is denoted as $V^i_j = \left\{\Delta^i_{j,k} | 1 \leqslant k \leqslant |K_i|\right\}$, and $\alpha^i_j(k)$ is used to indicate whether the binarized differences $\Delta^i_{j,k}$ is retained. If $\Delta^i_{j,k}$ is retained, $\alpha^i_j(k) = 1$. Otherwise, $\alpha^i_j(k) = 0$. Firstly, the continuous area with the value of 1 in the binarized difference are treated as particles, and those particles with the size greater than the average of all particle sizes are chosen to describe the binarized difference. Next, a quadruples, $\rho^i_{j,k} =< \rho^i_{j,k,1}, \rho^i_{j,k,2}, \rho^i_{j,k,3}, \rho^i_{j,k,4} >$ is constructed to characterize $\Delta^i_{j,k}$. Wherein $\rho^i_{j,k,1}$ is the sum of the size of all selected particles, $\rho^i_{j,k,2}$ is

the maximum ratio of Euler number to area of all selected particles; $\rho^i_{j,k,3}$ is the smallest ratio of area to the smallest circumscribed rectangle area of all selected particles, and $\rho^i_{j,k,4}$ is the smallest roundness of all selected particles. Then, the inappropriate binarized differences in V^i_j are cleaned according to their corresponding 4-tuple. The rule to eliminate the abnormal difference is as follows: If there is a frame difference, as long as its any one of properties (**the sum of the size, the maximum ratio of Euler number to area, the smallest ratio of area to the smallest circumscribed rectangle area, and the smallest roundness**) is an outlier in the set of corresponding property of all $\Delta^i_{j,k}(\in V^i_j)$, the difference will be eliminated. In addition, the threshold values used to judge whether the component values are abnormal is determined by Pauta Criterion. The specific thresholds used for 4 kinds of properties are $\tau_1 = \mu_1 - \frac{3}{4}\delta_1$, $\tau_2 = \mu_2 + \delta_2$, $\tau_3 = \mu_3 - \delta_3$ and $\tau_4 = \mu_4 - \delta_4$, respectively, wherein, μ_1 is the average of all $\rho^i_{j,k,1}$, δ_1 is the standard deviation all $\rho^i_{j,k,1}$, and so on. Subsequently, if the binarized differences $\Delta^i_{j,k}$, which 4 property values satisfy the conditions that first property value is greater than τ_1, the second property value is less than τ_2, the third property value is greater than τ_3, and the fourth property value is greater than τ_4 simultaneously, would be retained to detect the objects from the background. Otherwise, those differences that do not satisfy the conditions are discarded. As a result, a square matrix, $\alpha^i = \left[\alpha^i_1, \alpha^i_2, \ldots, \alpha^i_{|K_i|}\right]$, is obtained, where α^i_j is a column vector, defined as $\alpha^i_j = \left[\alpha^i_j(1), \alpha^i_j(2), \ldots, \alpha^i_j(|K_i|)\right]'$ wherein symbol $'$ denotes transpose operation. Finally, because difference $\Delta^i_{j,k}$ and $\Delta^i_{k,j}$ are the same, the selected state of them should be the same. Thus, $\alpha^i_j(k) = \alpha^i_k(j) = \alpha^i_j(k)\alpha^i_k(j)$ is used to represent whether $\Delta^i_{j,k}$ is retained.

D. GENERATING THE TEMPORARY BACKGROUND AND MASK

Assuming that the initial background reconstructed by the *jth* keyframe in K_i is noted as B^i_j, the background reconstructed by all keyframes in K_i is noted as B^i, and the temporary background reconstructed by all subsequence K_i is noted as B. To represent the mask for B^i_j, B^i and B respectively, the symbols, M^i_j, M^i and M are employed.

First of all, the retained binarized differences in α^i_j and the corresponding keyframes are employed to generate the background B^i_j. The gray value of B^i_j at the location (x, y) is defined as

$$B^i_j(x,y) = \sum_{k=1}^{|K_i|} H^i_{j,k} \cdot Z^i_{j,k}(x,y) \Big/ \sum_{k=1}^{|K_i|} H^i_{j,k}(x,y) \qquad (6)$$

under the condition that $\sum_{k=1}^{|K_i|} H^i_{j,k}(x,y) \neq 0$. Here, $H^i_{j,k}(x,y) = \sum_{k=1}^{|K_i|} \alpha^i_j(k)\overline{\Delta^i_{j,k}}(x,y)$, and $Z^i_{j,k}(x,y) = \frac{f_{i+\lambda(j-1)}(x,y)+f_{i+\lambda(k-1)}(x,y)}{2}$. $\alpha^i_j(k)$ indicates whether the binarized difference between *jth* keyframe and *kth* keyframe is selected, $\overline{\Delta^i_{j,k}}(x,y)$ is the inverted value of $\Delta^i_{j,k}(x,y)$ that indicates whether the pixel of the *jth* keyframe and *kth* keyframe are background at position(x, y), and $Z^i_{j,k}(x,y)$ is the average gray value of *jth* keyframe and *kth* keyframe at (x, y) in subsequence K_i. Simultaneously, the value of mask M^i_j at the location (x, y) is set to 1, if the value of $\sum_{k=1}^{|K_i|} H^i_{j,k}(x,y)$ is equal to 0. Otherwise, it is set to 0. In the event of $\sum_{k=1}^{|K_i|} H^i_{j,k}(x,y) = 0$, it signifies that all $\alpha^i_j(k)\overline{\Delta^i_{j,k}}$ values are 0, indicating that either the binarized difference is not selected ($\alpha^i_j(k) = 0$) or the pixel position represents the foreground target ($\overline{\Delta^i_{j,k}} = 0$). In this case, the background here cannot be determined at the moment and further processing is needed in the next stage. Then, the gray value of B^i at location (x, y) is defined as

$$B^i(x,y) = \sum_{j=1}^{|K_i|} \overline{M^i_j}(x,y) B^i_j(x,y) \Big/ \sum_{j=1}^{|K_i|} \overline{M^i_j}(x,y) \qquad (7)$$

under the condition that $\sum_{j=1}^{|K_i|} \overline{M^i_j}(x,y) \neq 0$, where $\overline{\ }$ represents the inverse operation. Meanwhile, the value of mask, $M^i(x,y)$ for B^i at location (x, y) is set to 1 if the value of $\sum_{j=1}^{|K_i|} \overline{M^i_j}(x,y)$ is equal to 0. Lastly, the gray value of B at location (x, y) is given by

$$B(x,y) = \sum_{i=1}^{\lambda} \overline{M^i}(x,y) B^i(x,y) \Big/ \sum_{i=1}^{\lambda} \overline{M^i}(x,y) \qquad (8)$$

on the case that $\sum_{i=1}^{\lambda} \overline{M^i}(x,y) \neq 0$. Additionally, the value of mask, $M(x, y)$, for B at location (x, y) is set to 1 if the value of $\sum_{i=1}^{\lambda} \overline{M^i}(x,y)$ is equal to 0. Otherwise, it is set to 0.

E. UPDATING THE TEMPORARY BACKGROUND

In this step, the estimation of the background beneath the masks is conducted. The proposed scheme for background estimation under the mask encompasses multiple sequential steps. Firstly, the mask matrix M is partitioned into smaller blocks of size $\varphi(= \frac{wh}{2^4 \times 2^4})$. Subsequently, a comparison is performed between A_i, representing the area of the mask, and 4φ. If the area of the mask is below 4φ, the process of selecting frames to reconstruct the background under the mask is initiated. However, if the mask's area exceeds 4φ, an iterative process is employed to reconstruct the background under the individual blocks. Priority is given to the blocks located near the boundary for background reconstruction, prioritizing them over the other blocks until all blocks within the mask have been processed. Finally, if there are no remaining masks to be processed, the temporary background serves as the final estimated background. The scheme of updating temporary background is depicted in Fig.2. The key aspects will be elaborated in detail below.

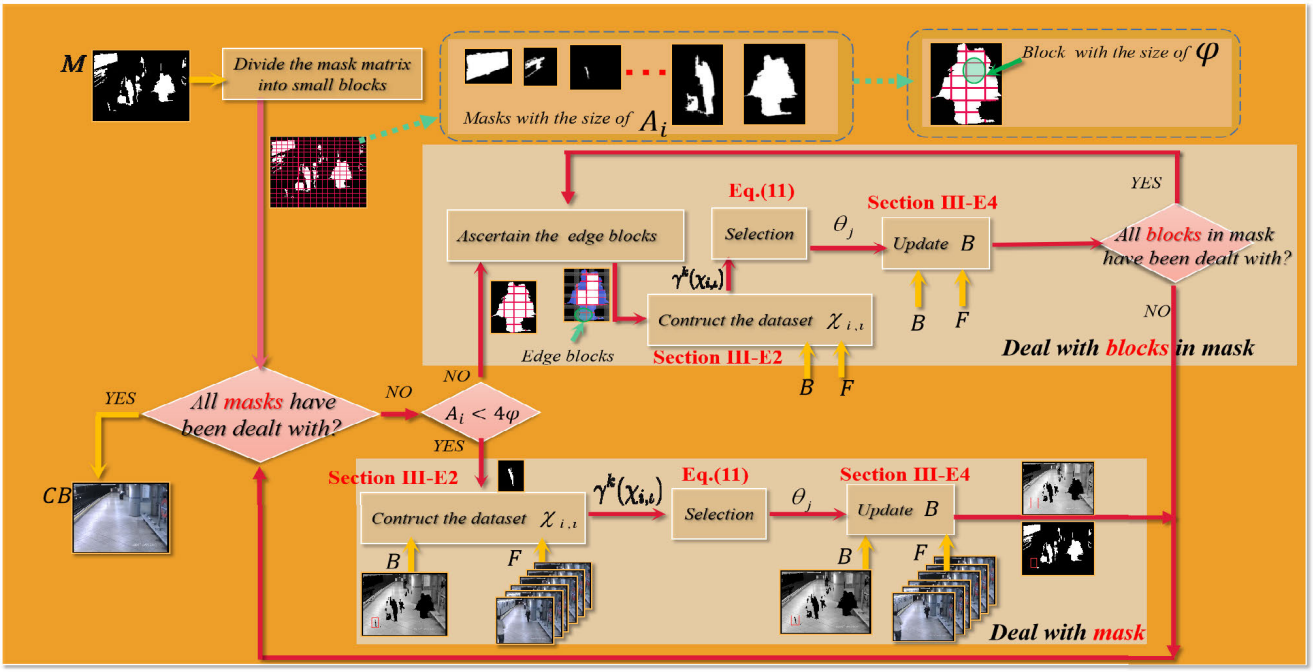


FIGURE 2. The scheme for updating the temporary background. Firstly, the mask matrix is divided into small blocks. Next, a comparison is made between the mask's area, A_i, and the threshold, 4φ. If the area of the mask is below the threshold, frames are selected to reconstruct the background under the mask. Otherwise, if the mask's area exceeds the threshold, frames are iteratively selected to reconstruct the background under the blocks located within the boundary until all blocks of this mask have been processed. Finally, if there are no remaining masks to be processed, the current temporary background serves as the final reconstructed background (or Computed Background, CB). Otherwise, the temporary background will continue to be updated.

1) LOCATIONS FOR SIMILARITY COMPARISON

As shown in Fig.3, for the mask in M, there exist several regions noted by different colors. These regions consist of the mask boundary(■), the adjacent exterior area(■), the adjacent interior area(■), the mask center(■), and the interior area of the mask(■). Assuming that *ith* mask in M is noted as M_i, the locations of pixels in the boundary of M_i is noted as L_i^b, the locations of interior pixels of M_i is noted as L_i, the locations of pixels in the external area adjacent to the boundary of M_i is defined as $L_i^e = \{(x,y)|\,|x-x'| \leqslant 3, |y-y'| \leqslant 3, M(x,y) == 0\}$, where $(x',y') \in L_i^b$, conditions that $|x-x'| \leqslant 3$ *and* $|y-y'| \leqslant 3$ illustrate L_i^e with a distance of less than 3 from the boundary. Moreover, $M(x,y) == 0$ implies that L_i^e is at the outside of the boundary. Correspondingly, the location set of interior pixels adjacent to the boundary is defined as $L_i^i = \{(x,y)|\,|x-x'| \leqslant 3, |y-y'| \leqslant 3, M(x,y) == 1\}$, where $(x',y') \in L_i^b$.

2) DATA USED TO COMPARE

Several data sets are suggested for the purpose of assessing the local similarity between B and each of input frames (See Fig. 3). For the *ith* mask (or block in mask), the first data set, denoted as $\chi_{i,1}$, where each element represents the difference between the average gray values of B at L_i^e and the corresponding values in each input frame at L_i^i. The second data set, denoted as $\chi_{i,2}$, comprises elements that are the difference between the standard deviation of gray values of B at L_i^e and the corresponding values in each input frame at L_i^i. The third data set, denoted as $\chi_{i,3}$, consists of individual elements that are the mean absolute errors (*MAE*) produced by the gray values of B and the corresponding values in each input frame at L_i^e. The *MAE* can be obtained by Eq.(9)

$$MAE_i(j) = \sum_{(x,y)\in L_i^e} |f_j(x,y) - B(x,y)| \Big/ |L_i^e| \tag{9}$$

where $|L_i^e|$ denotes the number of pixels at L_i^e. $MAE_i(j)$ reflects the deviation of input frame f_i from B at the external area adjacent the boundary. The fourth data set, denoted as $\chi_{i,4}$, comprises elements that are the root mean square errors (*RMSE*) produced by gray values of B and the corresponding values in each input frame at L_i^e. It can be obtained by Eq.(10).

$$RMSE_i(j) = \sqrt{\sum_{(x,y)\in L_i^e} (f_j(x,y) - B(x,y))^2 \Big/ |L_i^e|} \tag{10}$$

The fifth data set, denoted as $\chi_{i,5}$, comprises elements that are the differences between the mean of gray values of B at L_i^e and the corresponding values in each input frame at L_i. The sixth data set, denoted as $\chi_{i,6}$, comprises elements that are the difference between the mean of gray values of B at L_i^e and the value at the center of each input frame. The scheme of using six datasets constructed with local information to compare local similarity is more resilient in complex environments.

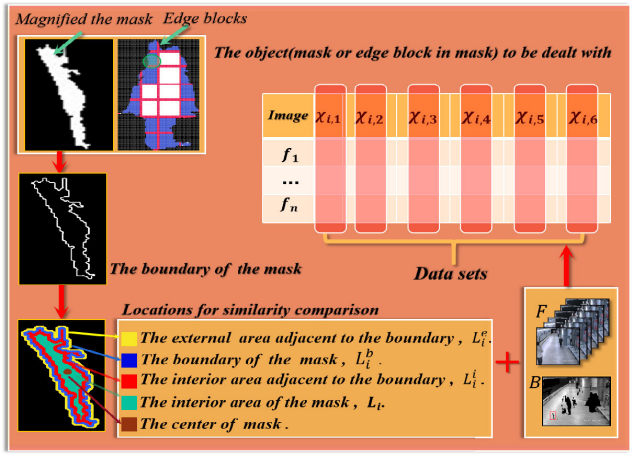


FIGURE 3. The locations and the compiled datasets utilized for the purpose of conducting similarity comparison. (■ represents the boundary of the mask or edge block, denoted as L_i^b. ■ represents the external area adjacent to the boundary of the mask or edge block, denoted as L_i^e. ■ represents the interior area adjacent to the boundary of the mask or edge block, denoted as L_i^i. ■ represents the interior area of the mask or edge block, denoted as L_i. ■ represents the center of the mask or edge block. $\chi_{i,1}$ is the set of the difference between the average gray values of B at L_i^e and the corresponding values in each input frame at L_i^i. $\chi_{i,2}$ is the set of the difference between the standard deviation of gray values of B at L_i^e and the corresponding values in each input frame at L_i^i. $\chi_{i,3}$ is the set of *MAE*, produced Eq.(9). $\chi_{i,4}$ is the set of *RMSE*, produced by Eq.(10). $\chi_{i,5}$ is the set of the differences between the mean of gray values of B at L_i^e and the corresponding values in each frame at L_i. $\chi_{i,6}$ is the set of the difference between the mean of gray values of B at L_i^e and the value at the center of each input frame.)

3) THE PROCESS USED TO SELECT CANDIDATE FRAMES

Having obtained crucial data from proposed locations, which are used to judge the similarity between frames and the known background, the Pauta Criterion (Eq.(11)) was employed to screen these data and select appropriate frames according to the screening results in order to generate the background under masks. In the updating temporary background stage, θ_j, with an initial value of 0, is used to represent whether the *jth* frame is reserved to reconstruct the background. If the *jth* frame is selected, θ_j is set to 1. The key process is as follows, accompanied by a few noteworthy considerations. Firstly, the selection process should be carried out in a certain order. Based on experiments, it is highly recommended to consider the execution sequences of $\chi_{i,1}$, $\chi_{i,2}$, $\chi_{i,3}$, $\chi_{i,4}$, $\chi_{i,5}$ and $\chi_{i,6}$ twice. Secondly, the threshold, γ^k, used to select appropriate candidate frames for the *kth* round selection is set according to Eq.(11)

$$\gamma^k(\chi_{i,\iota}) = \mu(\chi_{i,\iota}) + \phi_\iota^k \delta(\chi_{i,\iota}) \tag{11}$$

where, i represents the sequence number of mask; $\iota(= 1, 2, 3, 4, 5, 6)$ represents the category of χ to be dealt with; k(=1,2) represents the round of selection; $\mu(\cdot)$ represents the mean function; $\delta(\cdot)$ represents the standard deviation function; and ϕ_ι^k is a constant related to the category of $\chi_{i,\iota}$ and the round of selection. Specially, only all the elements in six data sets related to frame f_j are less than or equal to $\gamma^k(\chi_{i,\iota})$, the frame f_j is retain to reconstruct the background under the *ith* mask and θ_j is set to 1.

4) UPDATING THE TEMPORARY BACKGROUND

At this stage, the background beneath the blocks or mask is estimated using the selected frames according to Eq. (12).

$$B(x, y) = \overline{M_i(x, y)} \times B(x, y) + M_i(x, y) \times \sum_{j=1}^{n} \theta_j \times f_j(x, y) \Big/ \sum_{j=1}^{n} \theta_j \tag{12}$$

wherein, the first part is derived from the temporary background, and the second part is updated by the selected candidate frames.

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

1) DATASETS

The model was implemented on two distinct datasets, namely the SBMnet dataset [34][1] and the SBI dataset [20].[2] The SBMnet dataset was chosen for its ability to provide a diverse set of videos that cover a wide range of challenges in scene background modeling. This dataset is representative of typical indoor and outdoor visual data captured in surveillance, smart environment, and video database scenarios. It comprises 79 videos, including those from public datasets such as CDnet, LIMU, CMU, ATON, Fish4Knowledge, UCF, and MIT, among others. It is a realistic and diverse dataset that presents various challenges, including Basic, Intermittent Motion, Clutter, Jitter, Illumination Changes, Background Motion, Very Long, and Very Short. Conversely, the SBI dataset contains a sizable amount of data extracted from publicly available sequences.

2) EVALUATION METRICS

The performance of the proposed model was evaluated using six error measures:

AGE -Average of the gray-level absolute difference between groundtruth (GT) and the final background image. The final background image is alternatively referred to as the computed background (*CB*).

pEPs-Percentage of EPs (number of pixels in *CB* whose value differs from the value of the corresponding pixel in *GT* by more than a threshold) with respect to the total number of pixels in the image.

pCEPs-Percentage of CEPs (number of pixels whose 4-connected neighbors are also error pixels) with respect to the total number of pixels in the image.

MSSSIM-MultiScale Structural Similarity Index, an estimation of the perceived visual distortion.

PSNR-Amounts to $10log_10((L-1)^2/MSE)$ where L is the maximum number of gray levels and MSE is the Mean Squared Error between *GT* and *CB* images.

CQM-Color image Quality Measure with values expressed in decibels, where the higher the CQM value is, the better the background is.

[1]http://scenebackgroundmodeling.net/, accessed on 22 January 2022

[2]https://sbmi2015.na.icar.cnr.it/SBIdataset.html, accessed on 22 January 2022

3) IMPLEMENTATION DETAILS

A desktop computer equipped with an Intel(R) Core(TM) i9-9900 CPU@3.10GHz, but without a dedicated GPU, was utilized to experiments. The model implementation was carried out using Matlab R2019b. Several system parameters needed to be configured within the model, including:

$\boldsymbol{F}$, the input frame sequence.

$\boldsymbol{\zeta}$, the parameters employed for regulating the queue size in the process of extracting keyframes.

$\boldsymbol{\lambda}$, the fixed interval employed for partitioning keyframe sequences into subsequences.

$\boldsymbol{\phi_\iota^k}$, coefficients utilized to calculate the thresholds for $\chi i, \iota$.

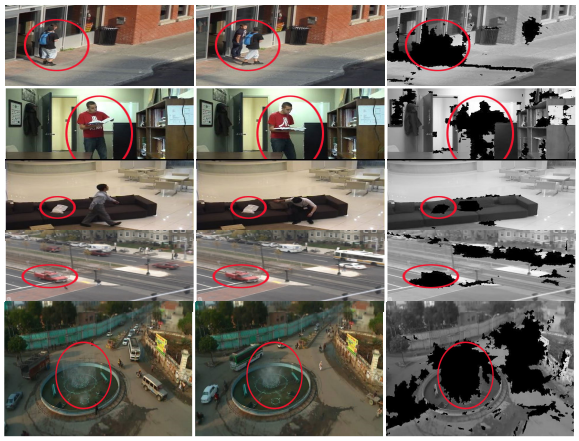
To summarize, the model required the configuration of the input frame sequence, F, as well as the parameters ζ (controlling the number of extracted keyframes), λ (sampling interval), and the coefficient $\chi_{i,\iota}$ for each filtering round. For the reported experiments, ζ was consistently set to 20, as the Pauta Criterion specifies a minimum of 10 samples. Considering that execution times tend to increase with a larger number of samples, λ was set to 10 for all reported experiments. During the first selection, the values of ϕ_1^1, ϕ_2^1, ϕ_3^1, ϕ_4^1, ϕ_5^1 and ϕ_6^1 were assigned as -0.75, 1, 0, 0.5, 1, and 0.25, respectively. Subsequently, for the second selection, these values were adjusted to 1, 0.25, 0.75, 1, 1, and 1, respectively, based on a significant number of repeated experiments.

B. QUALITATIVE RESULTS

The present study reports on the experimental outcomes of employing the proposed model for reconstructing the background across all videos included in the SBMnet and SBI datasets. The efficacy of the proposed methods is demonstrated in this section through the presentation of both the initial and final backgrounds. Specifically, the temporary backgrounds for the processed videos are displayed in the third column of Fig. 4, while the reconstructed backgrounds obtained via the proposed model are presented in the last row of Fig. 5, and Fig. 6.

1) THE TEMPORARY BACKGROUND

Fig. 4 showcases selected temporary backgrounds generated by the model, accompanied by corresponding masks. In Fig. 4, each row represents a video sequence, namely "busStation," "office," "sofa," "Uturn," and "IndianTraffic3." For the "busStation" video, consisting of a total of 617 frames, the foreground objects remain stationary in the same position for nearly the first 500 frames. In the "office" video, comprising 1449 frames, a man wearing a red shirt remains in the same position for approximately 1300 frames in the middle of the sequence. In the "sofa" video, consisting of 2600 frames, a white bag remains on the sofa for nearly 1700 frames in the middle of the sequence. In the "Uturn" video, comprising 479 frames, a small red car waiting to make a U-turn remains in the same position for almost 380 frames. The "IndianTraffic3" video, consisting of 901 frames, has approximately 320 frames in the middle where a fountain keeps spraying water. In Fig. 4, the first two columns display two frames from each of these videos. It can be observed that across different frames, the foreground objects remain nearly stationary without any significant movement. Other algorithms often struggle to differentiate these almost static foreground objects from the background. However, the proposed model effectively addresses this challenge by employing the aforementioned measures of "keyframe extraction", "grouping", and "the use of 4-tuple with particle shape attributes to filter binarized differences". These measures successfully separate the nearly static foreground objects from the background. From the third column of Fig. 4, it can be seen that during the generation of temporary backgrounds, the masks effectively distinguish these regions and cover them. Subsequently, the background generation for these regions is carried out in the update stage. This highlights the effectiveness of the strategies employed in this paper, namely keyframe extraction, grouping, and the use of 4-tuple with particle shape attributes, and handling regions where foreground objects and background are prone to confusion in later stages.



(a) Random frame 1 (b) Random frame 2 (c) Object mask

FIGURE 4. The temporary background with occlusion mask for long-term stationary object regions. (The first and second column is two random frames where foreground objects almost keep still. The third column is the temporary background with the mask.)

2) THE RECONSTRUCTED BACKGROUND

The proposed model was subjected to comprehensive experimentation involving 79 videos and the reconstructed backgrounds have been uploaded to the SBMnet website. The results of proposed model and the other state-of-the-art methods for SBMnet videos are presented in Fig. 5. It can be seen that, in the red rectangle of the first column, the reconstructed backgrounds for AVSS2007 video obtained from the methods of BB-SGD [16], LaBGen-P-Semantic (CV+U) [23], SPMD [24] or RMR [25] either blend in with the women sitting in benches, or mix up the train which stays for a long duration. In the second column,

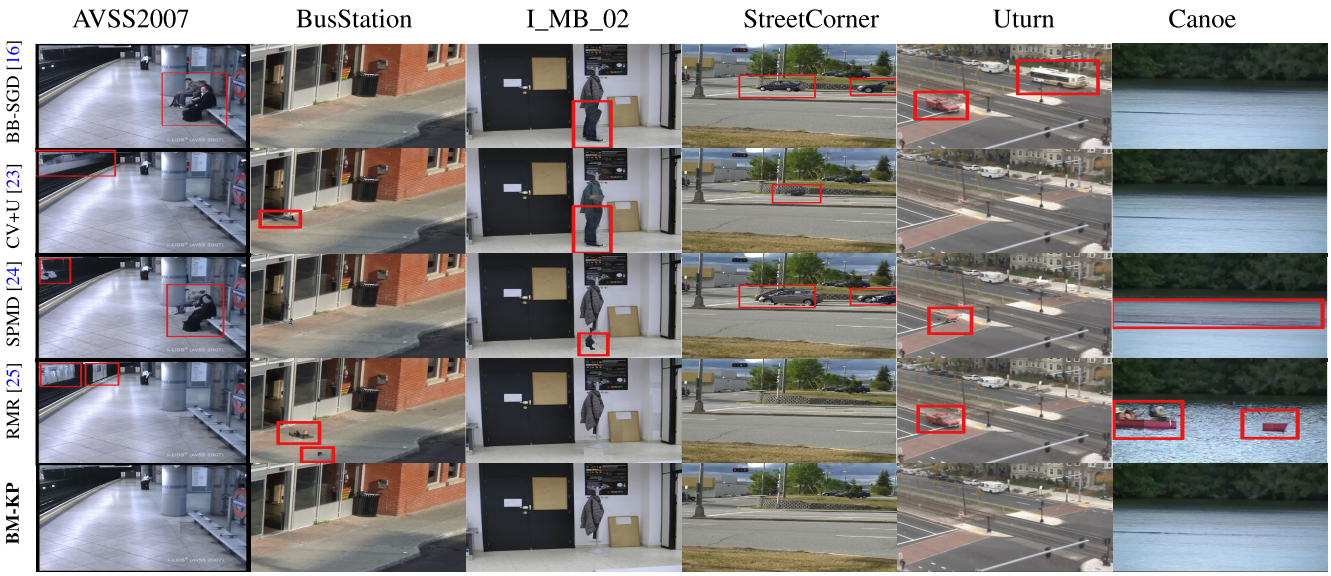


FIGURE 5. Some reconstructed backgrounds of our model and the other typical state-of-the-art methods for videos of SBMnet dataset. The first row is the result of BB-SGD [16] method for AVSS2007, BusStation, I_MB_02, StreetCorner, Uturn and Canoe video. The second row is the result of LaBGen-P-Semantic (CV+U) [23] for the six videos, the third row is the result of SPMD [24] method, the fourth row is the result of RMR [25], and the last row is the result of the proposed model repectively.

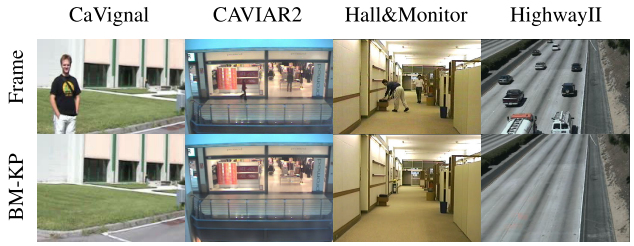


FIGURE 6. Reconstructed backgrounds using the proposed model for some videos of SBI dataset (The first row is example frames, the second row is the reconstructed background by the proposed model).

for the busstation video, the result of paper [23] blends the background with the shadow, and the result of [25] mixes the estimated background with the shoes. For the reconstructed background of I_MB_02, the first three methods fail to distinguish trousers from clothes hanging on the hanger, mistakenly taking trousers as the background. In the last three columns, the results in the red rectangle also demonstrate that the other three methods are not robust in distinguishing between the real background and the stationary objects. However, the reconstructed backgrounds depicted in the final row of Fig.5 indicate that the proposed model has achieved favorable results for the listed videos. This is because the proposed algorithm, after identifying the regions where nearly stationary foreground objects persist in the previous stage, employs a distinct background generation strategy for these regions in the second stage. Specifically, it determines the candidate frames for background generation by comparing the similarity between the background surrounding the mask in the temporary background and the corresponding areas in the input frame. One of the main reasons why other algorithms fail to accurately differentiate between foreground and background is that they adopt the same strategy for both the regions where nearly stationary foreground objects reside and other areas. Consequently, these algorithms are prone to erroneously treating the foreground as part of the background, particularly in videos with long-term stationary foreground objects. The results depicted in Fig. 5 confirm the efficacy of the background generation strategy employed by the proposed model for regions located below the mask during the updating stage. Furthermore, Fig. 6 presents the results of the proposed model on the SBI dataset, further confirming its robustness across different scenarios.

C. QUANTITATIVE EVALUATION AND COMPARISON

The quantitative evaluation of the proposed model and comparison with the other methods on SBMnet dataset and SBI datasets are presented in Table 3-8 and Fig. 7. The quantitative results presented in Table 3 demonstrate that the proposed algorithm achieves the best performance among the six metrics on the AVSS2007 videos. This is

TABLE 3. Performance comparison between the proposed model and the other state-of-the-art methods for AVSS2007 video(↓indicates lower score is better, ↑ indicates higher score is better.)

Method	AGE↓	pEPs↓	pCEPS↓	MSSSIM↑	PSNR↑	CQM ↑
MSCL [9]	7.5256	0.0612	0.0458	0.9294	22.3138	23.0990
BB-SGD [16]	8.73	N/A	N/A	N/A	N/A	N/A
FC-FlowNet [17]	11.6751	0.1214	0.0966	0.8726	20.7442	21.7565
LaBGen(CV+U) [23]	7.1867	0.0628	0.0490	0.9330	23.5443	24.2257
SPMD [24]	9.2676	0.0762	0.0544	0.8818	21.5274	22.4565
RMR [25]	9.2767	0.0663	0.0513	0.9094	20.3096	21.3404
Q-DMD [26]	17.3175	0.2589	0.2031	0.7123	19.3450	20.2835
DCP [27]	7.3008	N/A	N/A	N/A	N/A	N/A
BACV [29]	10.8909	0.0832	0.0491	0.8490	21.0333	21.9987
NExBI [14]	12.3242	0.0947	0.0697	0.8799	21.1518	22.0076
LaBGen [31]	8.3062	0.0707	0.0541	0.9050	21.4577	22.3158
BEWiS [32]	17.0903	0.1382	0.1128	0.7784	17.0107	18.0036
BM-KP(Ours)	**5.6437**	**0.0224**	**0.0131**	0.9628	28.5347	29.1752

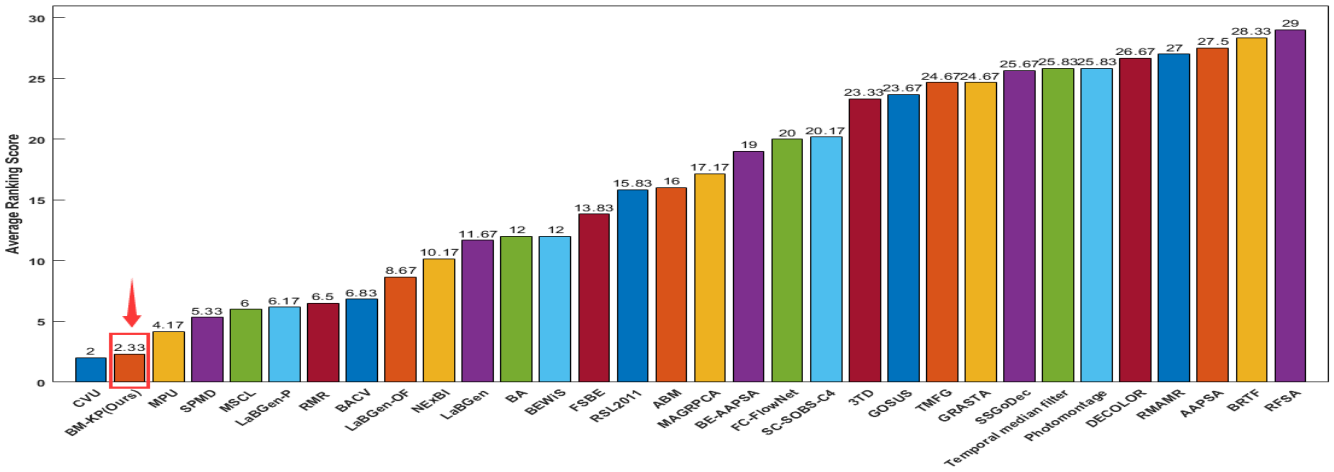


FIGURE 7. The average ranking score of the proposed model and the other 31 state-of-the-art methods for intermittent category video. The proposed model obtains 2.33 score(the lower score, the better), ranking the second, the first method obtains 2 score and the third method obtains 4.17 score.

attributed to the inability of other algorithms to effectively generate backgrounds for the regions in the video that contain long-term seated passengers and departing subway trains after a certain period. In contrast, the proposed model excels in generating backgrounds for these specific regions. Table 4 presents the average scores for the intermittent category of SBMnet and demonstrates that the proposed model outperforms all referenced methods in terms of average pCEPS, average MSSSIM, average PSNR, and CQM results. However, it falls short in terms of the average AGE and average pEPs metrics compared to the LaBGen(CV+U) [23]. This is attributed to the fact that in the first stage of the proposed model, when generating the background excluding the regions where stationary foreground objects reside, only a subset of keyframes is utilized. Consequently, these areas of the generated background rely on limited information from a small number of input frames, rather than utilizing information from all frames. This leads to higher values for the average AGE and average pEPs metrics for the proposed model.

TABLE 4. Performance comparison between the proposed model and the other state-of-the-art methods for intermittent motion videos.

Method	Average AGE↓	Average pEPs↓	Average pCEPS↓	Average MSSSIM↑	Average PSNR↑	CQM ↑
MSCL [9]	3.9743	0.0313	0.0215	0.9831	32.6916	33.4541
BB-SGD [16]	4.8898	0.0298	0.0171	0.9638	30.3043	31.1483
FC-FlowNet [17]	6.7811	0.0599	0.0347	0.9312	27.0272	27.9086
LaBGen(CV+U) [23]	**3.8079**	**0.0194**	0.0082	0.9810	32.8806	33.5830
SPMD [24]	4.1840	0.0206	0.0088	0.9745	31.9703	32.7043
RMR [25]	4.3606	0.0213	0.0091	0.9730	31.1372	31.9285
DCP [27]	5.3666	N/A	N/A	N/A	N/A	N/A
LaBGen-P [28]	4.1278	0.0225	0.0115	0.9712	31.9974	32.726
BACV [29]	4.5966	0.0198	0.0079	0.9657	30.2170	31.1074
NExBI [14]	4.6374	0.0248	0.0123	0.9639	30.0986	30.8849
BEWiS [32]	4.7798	0.0277	0.0173	0.9585	29.7747	30.6778
FSBE [33]	5.3438	0.0399	0.0220	0.9610	29.1968	30.1392
BM-KP (Ours)	4.1284	0.0208	**0.0074**	0.9842	32.8955	33.6616

To further evaluate the method's performance, a comparative analysis was conducted between the proposed model and state-of-the-art deep learning methods that were accessible for data acquisition. Table 5 presents the comparison in AGE measure between the proposed method and two deep learning methods on the ''background motion'' and ''illumination changes'' categories of SBMnet dataset. From the table, it can be seen that in some videos, the performance of the proposed method is superior to DCP [27] method and BI-GAN [18] method, which supports the effectiveness of the proposed method.

TABLE 5. Comparison in AGE measure between the proposed method and two deep learning methods on "background motion" and "illumination changes" categories of SBMnet dataset. The best AGE scores for each video sequence is shown in black.

Category	Videos	DCP [27]	BI-GAN [18]	**BM-KP**
Background motion	Canoe	**6.3250**	11.0422	12.7549
	Advertisement board	2.3378	4.5574	**1.8443**
	Fall	19.0737	**8.8745**	23.5028
	Fountain 01	9.6775	8.7117	**6.6859**
	Fountain 02	14.0579	**4.0120**	6.6196
	Overpass	**6.4089**	8.9543	7.9433
Illumination changes	Camera parameter	6.2206	7.4353	**4.3716**
	Dataset3 camera1	14.5708	**5.9966**	6.3202
	Dataset3 camera2	18.7047	5.5740	**4.7595**
	I_IL_ 01	**7.4329**	11.2693	7.8426
	I_IL_ 02	19.3833	**9.9795**	10.1225
	Cubicle	11.4636	**3.8472**	5.0864

Fig. 7 presents the average ranking, which is computed based on six metrics, including the proposed model and 31 other representative methods for the intermittent category on the SBMnet dataset (a dataset for testing background estimation algorithms). Our model achieves a score of 2.33 (lower scores indicating better performance), which is only 0.33 higher than the LaBGen(CV+U) [23] method and 1.84 lower than the third-ranked method. This ranking comprehensively reflects the commendable performance of the proposed model across the six metrics when dealing with intermittent motion videos.

The previous comparative analysis has already demonstrated the advanced performance of the proposed model in handling intermittent motion videos. To further validate the robustness of the proposed model, Table 6 presents the metrics obtained by the proposed model on the SBMnet dataset for all types of videos, including Basic, Intermittent Motion, Clutter, Jitter, Illumination Changes, Background Motion, Very Long, and Very Short. A comparison is made with 12 other state-of-the-art algorithms. From Table 6, it can be observed that although the proposed model performs lower than algorithms like BB-SGD [16] in terms of the six metrics, it outperforms FC-FlowNet [17] and RMR [25] algorithms. Additionally, it surpasses BACV [29] and LaBGen(MP+U) [23] algorithms in certain performance indicators. The reason for the outstanding performance of the proposed model in intermittent motion videos, but relatively lower overall evaluation, mainly lies in its suboptimal performance in handling "Jitter" and "Clutter" videos. Tables 3, 4, 5, and 6 collectively demonstrate that the proposed model excels in handling intermittent videos and exhibits a certain degree of robustness in complex scenes.

TABLE 6. Performance comparison between the proposed model and the other state-of-the-art methods for all videos of SBMnet dataset.

Method	Average AGE↓	Average pEPs↓	Average pCEPS↓	Average MSSSIM↑	Average PSNR↑	CQM ↑
MSCL [9]	5.9547	0.0524	0.0171	0.9410	30.8952	31.7049
BB-SGD [16]	5.6266	0.0447	0.0147	0.9478	30.4016	31.2420
FC-FlowNet [17]	9.1131	0.1128	0.0599	0.9162	26.9559	27.8767
LaBGen(CV+U) [23]	7.3890	0.0761	0.0357	0.9267	28.5050	29.3829
SPMD [24]	6.0985	0.0487	0.0154	0.9412	29.8439	30.6499
RMR [25]	9.5363	0.1176	0.0582	0.8790	26.5217	27.4549
BACV [29]	8.5816	0.0724	0.0257	0.9078	26.1018	27.1000
NExBI [14]	6.7778	0.0671	0.0227	0.9196	27.9944	28.8810
LaBGen [31]	6.7090	0.0631	0.0265	0.9266	28.6396	29.4668
BEWiS [32]	6.7094	0.0592	0.0266	0.9282	28.7728	29.6342
FSBE [33]	6.6204	0.0605	0.0217	0.9373	29.3378	30.1777
LaBGen(MP+U) [23]	7.9731	0.0820	0.0394	0.9212	28.3234	29.1992
BM-KP(Ours)	7.8418	0.0901	0.0424	0.9201	28.4173	29.3087

To provide a more comprehensive assessment of the capabilities of the proposed model in challenging scenarios, experiments were conducted using the SBI dataset. Table 7 shows that, although there are a few failure cases, such as 'Snellen', the proposed method is still effective for most videos of SBI dataset. The results presented in Table 8 illustrate that the performance of the proposed model on the SBI dataset is comparable to that of specific deep learning algorithms. Notably, the suggested model exhibited superior performance across various metrics, outperforming other methods in all videos except for HallAndMonitor, achieving victories in 26 out of 30 metrics across 6 videos. In contrast, BI-GAN [18] only achieved success in 4 out of 30 metrics.

D. COMPUTATIONAL ANALYSIS

Besides background image quality, the processing speed is also an important factor in evaluating the performance of the method. According to Fig. 1, the time cost of the proposed model can be broken down into three parts: extracting keyframe and grouping, computing binarized difference and screening, and generating the initial background and updating it. For the first part, in the worst case, the number of s_i with a local maximum or minimum will not exceed half of the total number. Thus, the complexity of this part is $O(\frac{n-\eta}{1-1/2})$, eg.$O(2(n-\eta))$. Assuming that the number of particles with a size greater than the average size of all particles is ξ. Thus, the complexity of the second part is $O(\lceil\frac{\eta}{\lambda}\rceil \times \frac{\lambda(\lambda-1)}{2} + \xi \times \lceil\frac{\eta}{\lambda}\rceil \times \frac{\lambda(\lambda-1)}{2} + 4 \times \lceil\frac{\eta}{\lambda}\rceil \times \lambda)$, eg. $O(\lceil\frac{\eta}{\lambda}\rceil \times (\frac{\lambda(\lambda-1)}{2}(1+\xi)+4\lambda))$. Moreover, $\lceil\frac{\eta}{\lambda}\rceil$ is fixed to 10 in the model. Therefore, the complexity of second part is $O(4.5\eta(1+\xi)+4\eta)$. Assuming that the total area of all particles with an area greater than 4φ is A. The complexity of the third part is $O(\lceil\frac{\eta}{\lambda}\rceil \times \lambda + \frac{A}{\varphi} \times 2 \times n)$, eg.$O(\eta + \frac{2nA}{\varphi})$. Taking this decomposition into account, the overall complexity of the proposed model, in the worst case, can be expressed as $O(2(n-\eta) + 4.5\eta(1+\xi) + 4\eta + \eta + \frac{2nA}{\varphi})$, i.e., the whole complexity is $O((7.5+4.5\xi)\eta + 2n(1+\frac{A}{\varphi}))$. The average speed of our method conducted on Matlab 2019b was 5.57 FPS (5.38 FPS for SBMnet dataset and 10.94 FPS for SBI dataset) with an Intel(R) Core(TM) i9-9900 CPU @ 3.10GHz without GPU. Notably, for the 'CaVignal' sequence, the speed was 35.68 FPS, which is faster than a real-time computing speed. Table 9 reports the comparison of processing speed for sequences with different resolutions, as well as the average speed for whole testing data, based on Frames Per Second (FPS) measure as reported publicly. ADNN [19] is a method based on the arithmetic distribution neural network, BScGAN is [38] a method based on generative adversarial network, BEWis [32] is a method based on weightless neural networks, and the others are traditional methods. It is necessary to be aware that the speed of the first three algorithms is the test speed, not including the training time. Some algorithms necessitate several hours or even days of training, thus the test speed alone is not sufficient to compare the performance of the algorithms. To the best of our knowledge, the fast background modeling is 52 FPS with GPU [16]. For traditional algorithms without GPU, the average computation speed of LaBGen-OF [30] is estimated to be 5 FPS in [30], which is lower than our method. The processing speed of SPMD [24] is estimated to

TABLE 7. Evaluation results of the proposed model on the SBI dataset.

Sequence	AGE	pEPs	pCEPS	MSSSIM	PSNR
Board	5.1915	1.4634	0.3232	0.9486	30.9897
Candela_m1.10	2.8876	0.2308	0.0000	0.9956	35.4889
CAVIAR1	2.4832	0.1689	0.0793	0.9948	37.4907
CAVIAR2	0.8517	0.0132	0.0031	0.9991	45.8774
CaVignal	1.8810	0.0184	0.0000	0.9957	39.0900
Foliage	17.3536	27.0590	18.4306	0.7301	18.6683
HallAndMonitor	2.5275	0.2344	0.0781	0.9853	34.0840
HighwayI	2.5011	0.2930	0.0404	0.9861	36.8374
HighwayII	2.1079	0.2174	0.0000	0.9941	37.4872
HumanBody2	4.6120	0.5143	0.0000	0.9950	32.4432
IBMtest2	3.5617	0.1445	0.0000	0.9898	34.7259
PeopleAndFoliage	18.9403	23.0495	19.4740	0.7044	16.0657
Snellen	25.3863	33.2948	29.3499	0.7634	15.7678
Toscana	6.0346	5.5606	4.4612	0.9140	24.2172

TABLE 8. A comparative analysis of the proposed model and BI-GAN [18] algorithm in Age, pEPs, pCEPS, and MSSSIM scores on six videos of SBI dataset. (The red indicators are the ones that the BI-GAN [18] algorithm wins, while the rest of the algorithms are the ones that the suggested model wins.)

Video	AGE↓		pEPs↓		pCEPS↓		MSSSIM↑		PSNR↑	
	BM-KP	BI-GAN [18]	**BM-KP**	BI-GAN [18]	**BM-KP**	BI-GAN [18]	**BM-KP**	BI-GAN [18]	**BM-KP**	BI-GAN [18]
Candela_m1.10	2.8876	4.9075	0.2308	0.9033	0.0000	0.0488	0.9956	0.9800	35.4889	31.3445
CAVIAR1	2.4832	8.5872	0.1689	7.0801	0.0793	0.8057	0.9948	0.9529	37.4907	26.2490
CAVIAR2	0.8517	12.7988	0.0132	13.6963	0.0031	2.0020	0.9991	0.9809	45.8774	24.8151
CaVignal	1.8810	4.5303	0.0184	2.1484	0.0000	0.0977	0.9957	0.9877	39.0900	30.8977
HallAndMonitor	2.5275	3.9255	0.2344	0.1709	0.0781	0.0000	0.9853	0.9899	34.0840	34.1297
HumanBody2	4.6120	7.1523	0.5143	4.1748	0.0000	0.2930	0.9950	0.9744	32.4432	28.7653

TABLE 9. Comparison of processing speed (for sequences with different resolutions and for whole testing data, Units: FPS).

Sequence	Foliage (200 × 144)	highway (320 × 240)	I_CA _01 (352 × 288)	wetSnow (536 × 320)	511 (640 × 480)	Average Speed
ADNN [19]	N/A	0.33	N/A	N/A	N/A	N/A
BScGAN [38]	N/A	N/A	N/A	N/A	N/A	10
BEWiS [32]	4.3	2.5	2.2	1.1	0.6	3
SPMD [24]	22.8	5.6	5.2	2.9	1.6	N/A
MSCL [9]	N/A	1.26	N/A	N/A	N/A	N/A
LaBGen-OF [30]	3.3	1.6	1.2	0.8	0.5	5
BM-KP	**24**	**12.2**	**11.5**	**4**	**1.39**	**5.57**

22.8 FPS, 5.6 FPS, 5.2 FPS, 2.9 FPS and 1.6 FPS for Foliage, highway, I_CA_01, wetSnow and 511, respectively. The time of MSCL [9] for the highway sequence is 39.8 seconds, eg, the speed is about 1.26 FPS. Additionally, the computational cost of 200 color 350×240 frames with average resolution of 240×349 is around 4.5 minutes for RMR [25], i e, the speed is about 1.35 FPS. The time of BACV [29] is about average 1.55376 ms per pixel, eg, for one 200×144 image, it's about 44.75 seconds (the speed of 0.02FPS). From the reported data, it can be verified that the proposed BM-KP method is faster than most methods. This further demonstrates the effectiveness of the keyframe extraction strategy.

V. CONCLUSION AND FUTURE WORK

The distinction between stationary objects and the background is a prominent subject in visual processing, and the associated technologies are highly demanded in video surveillance systems. In this paper, a model based on the keyframe and particle shape property is presented to address this issue. The model undergoes evaluation using the SBMnet dataset and the SBI dataset, followed by comparisons with current state-of-the-art methods. The results obtained indicate the effectiveness of the proposed model in distinguishing stationary objects from the background. The model demonstrates robustness in handling illumination changes and background motion, while also exhibiting faster performance compared to conventional methods. These findings from our study serve to demonstrate the efficacy and resilience of our method, thereby validating the proposed strategies.

Nevertheless, certain failure cases were observed in our method, including scenarios such as"Foliage", "PeopleAndFoliage", "Snellen" and "Toscana". Additionally, the thresholds utilized in the selection process were manually set based on iterative experimentation. In order to tackle these challenges, future investigations will explore the integration of unsupervised learning techniques into the proposed model.

REFERENCES

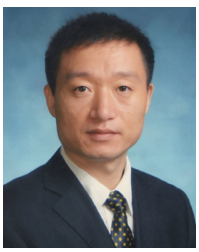
[1] H. Ren, L. Xing, and T. Shi, "Research on background learning correlation filtering algorithm with multi-feature fusion," *IEEE Access*, vol. 11, pp. 32895–32906, 2023.
[2] Y. Lu and S. Huang, "Sparse representation based hyperspectral anomaly detection via adaptive background sub-dictionaries," *IEEE Access*, vol. 9, pp. 14735–14751, 2021.
[3] S. Pei, L. Li, L. Ye, and Y. Dong, "A tensor foreground-background separation algorithm based on dynamic dictionary update and active contour detection," *IEEE Access*, vol. 8, pp. 88259–88272, 2020.
[4] J.-Y. Kim and J.-E. Ha, "Weakly supervised foreground object detection network using background model image," *IEEE Access*, vol. 10, pp. 105726–105733, 2022.
[5] C. Lyu, Y. Liu, X. Wang, Y. Chen, J. Jin, and J. Yang, "Visual early leakage detection for industrial surveillance environments," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 3670–3680, Jun. 2022.
[6] L. Ma, H. Qi, S. Zhu, and S. Ma, "A fast background model based surveillance video coding in HEVC," in *Proc. IEEE Vis. Commun. Image Process. Conf.*, Dec. 2014, pp. 237–240.
[7] G. Wang, B. Li, Y. Zhang, and J. Yang, "Background modeling and referencing for moving cameras-captured surveillance video coding in HEVC," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 2921–2934, Nov. 2018.
[8] M. O. Tezcan, P. Ishwar, and J. Konrad, "BSUV-Net: A fully-convolutional neural network for background subtraction of unseen videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2763–2772.
[9] S. Javed, A. Mahmood, T. Bouwmans, and S. K. Jung, "Background–foreground modeling based on spatiotemporal sparse subspace clustering," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5840–5854, Dec. 2017.
[10] A. Colombari and A. Fusiello, "Patch-based background initialization in heavily cluttered video," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 926–933, Apr. 2010.
[11] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359–373, Jan. 2015.
[12] B. Garcia-Garcia, T. Bouwmans, and A. J. R. Silva, "Background subtraction in real applications: Challenges, current models and future directions," *Comput. Sci. Rev.*, vol. 35, Feb. 2020, Art. no. 100204.
[13] W. Liu, Y. Cai, M. Zhang, H. Li, and H. Gu, "Scene background estimation based on temporal median filter with Gaussian filtering," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 132–136.
[14] W. S. Mseddi, M. Jmal, and R. Attia, "Real-time scene background initialization based on spatio-temporal neighborhood exploration," *Multimedia Tools Appl.*, vol. 78, no. 6, pp. 7289–7319, Mar. 2019.
[15] L. Li, Z. Wang, Q. Hu, and Y. Dong, "Adaptive nonconvex sparsity based background subtraction for intelligent video surveillance," *IEEE Trans. Ind. Informat.*, vol. 17, no. 6, pp. 4168–4178, Jun. 2021.
[16] B. Sauvalle and A. de La Fortelle, "Fast and accurate background reconstruction using background bootstrapping," *J. Imag.*, vol. 8, no. 1, p. 9, Jan. 2022.
[17] I. Halfaoui, F. Bouzaraa, and O. Urfalioglu, "CNN-based initial background estimation," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 101–106.

[18] M. Sultana, A. Mahmood, T. Bouwmans, and S. K. Jung, "Unsupervised adversarial learning for dynamic background modeling," in *Proc. Int. Workshop Frontiers Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 248–261.
[19] C. Zhao, K. Hu, and A. Basu, "Universal background subtraction based on arithmetic distribution neural network," *IEEE Trans. Image Process.*, vol. 31, pp. 2934–2949, 2022.
[20] T. Bouwmans, L. Maddalena, and A. Petrosino, "Scene background initialization: A taxonomy," *Pattern Recognit. Lett.*, vol. 96, pp. 3–11, Sep. 2017.
[21] K.-L. Hua, H.-C. Wang, C.-H. Yeh, W.-H. Cheng, and Y.-C. Lai, "Background extraction using random walk image fusion," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 423–435, Jan. 2018.
[22] Z. Bai, Q. Gao, and X. Yu, "Moving object detection based on adaptive loci frame difference method," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Aug. 2019, pp. 2218–2223.
[23] B. Laugraud, S. Piérard, and M. Van Droogenbroeck, "LaBGen-P-semantic: A first step for leveraging semantic segmentation in background generation," *J. Imag.*, vol. 4, no. 7, p. 86, Jun. 2018.
[24] Z. Xu, B. Min, and R. C. C. Cheung, "A robust background initialization algorithm with superpixel motion detection," *Signal Process., Image Commun.*, vol. 71, pp. 1–12, Feb. 2019.
[25] D. Ortego, J. C. SanMiguel, and J. M. Martínez, "Rejection based multipath reconstruction for background estimation in video sequences with stationary objects," *Comput. Vis. Image Understand.*, vol. 147, pp. 23–37, Jun. 2016.
[26] J. Han, K. I. Kou, and J. Miao, "Quaternion-based dynamic mode decomposition for background modeling in color videos," *Comput. Vis. Image Understand.*, vol. 224, Nov. 2022, Art. no. 103560.
[27] M. Sultana, A. Mahmood, S. Javed, and S. K. Jung, "Unsupervised deep context prediction for background estimation and foreground segmentation," *Mach. Vis. Appl.*, vol. 30, no. 3, pp. 375–395, Apr. 2019.
[28] B. Laugraud, S. Piérard, and M. Van Droogenbroeck, "LaBGen-P: A pixel-level stationary background generation method based on LaBGen," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 107–113.
[29] T. Minematsu, A. Shimada, and R.-I. Taniguchi, "Background initialization based on bidirectional analysis and consensus voting," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 126–131.
[30] B. Laugraud and M. Van Droogenbroeck, "Is a memoryless motion detection truly relevant for background generation with LaBGen?" in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.* Cham, Switzerland: Springer, 2017, pp. 443–454.
[31] B. Laugraud, S. Piérard, and M. Van Droogenbroeck, "LaBGen: A method based on motion detection for generating the background of a scene," *Pattern Recognit. Lett.*, vol. 96, pp. 12–21, Sep. 2017.
[32] M. De Gregorio and M. Giordano, "Background estimation by weightless neural networks," *Pattern Recognit. Lett.*, vol. 96, pp. 55–65, Sep. 2017.
[33] A. Djerida, Z. Zhao, and J. Zhao, "Robust background generation based on an effective frames selection method and an efficient background estimation procedure (FSBE)," *Signal Process., Image Commun.*, vol. 78, pp. 21–31, Oct. 2019.
[34] P.-M. Jodoin, L. Maddalena, A. Petrosino, and Y. Wang, "Extensive benchmark and survey of modeling methods for scene background initialization," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5244–5256, Nov. 2017.
[35] H. Liu, W. Meng, and Z. Liu, "Key frame extraction of online video based on optimized frame difference," in *Proc. 9th Int. Conf. Fuzzy Syst. Knowl. Discovery*, May 2012, pp. 1238–1242.
[36] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
[37] E. Berrezueta, J. Cuervas-Mons, Á. Rodríguez-Rey, and B. Ordóñez-Casado, "Representativity of 2D shape parameters for mineral particles in quantitative petrography," *Minerals*, vol. 9, no. 12, p. 768, Dec. 2019.
[38] M. C. Bakkay, H. A. Rashwan, H. Salmane, L. Khoudour, D. Puig, and Y. Ruichek, "BSCGAN: Deep background subtraction with conditional generative adversarial networks," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4018–4022.

YONG FAN received the B.S. degree in education technology from Southwest Normal University, in 2002, and the M.S. degree in computer application technology from Southwest Petroleum University, in 2014. He is currently pursuing the Ph.D. degree in computer technology and application with the School of Computer Science and Engineering, Macau University of Science and Technology, Taipa, Macau, China. His research interest includes image and video processing.

XIU HE received the B.S. degree from Xinzhou Teachers University, Xinzhou, China, in 2010, and the M.S. degree from the Taiyuan University of Technology, Taiyuan, China, in 2013. She is currently pursuing the Ph.D. degree with the Faculty of Information Technology, Macau University of Science and Technology, Macau, China. She is also a Lecturer with the School of Information Science, Xinhua College, Sun Yat-sen University, Guangzhou (Dongguan), China. Her research interests include signal and image processing and information retrieval.

YIYI LIN received the B.S. degree in geomatics engineering from South China Agricultural University, Guangzhou, China, in 2021. He is currently pursuing the M.S. degree in space big data analytics with the State Key Laboratory of Lunar and Planetary Sciences, Macau University of Science and Technology, Taipa, Macau, China. His research interests include image processing and remote sensing data processing and analysis.



ZHANCHUAN CAI (Senior Member, IEEE) received the Ph.D. degree in computer software and theory from Sun Yat-sen University, Guangzhou, China, in 2007. He is currently a Professor with the School of Computer Science and Engineering, Macau University of Science and Technology, Macau, China, where he is also with the State Key Laboratory of Lunar and Planetary Sciences. He has authored over 100 papers in refereed journals and conferences. His research interests include image processing and computer graphics, intelligent information processing, multimedia information security, and remote sensing data processing and analysis.