Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2021) 29: 152 – 168

doi:10.3906/elk-1904-187

Abnormal behavior detection using sparse representations through sequential generalization of *k*-means

Ahlam AL-DHAMARI[1,2,*], **Rubita SUDIRMAN**[1], **Nasrul Humaimi MAHMOOD**[1]

[1]Division of Electronic and Computer Engineering, School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, Johor, Malaysia

[2]Department of Computer Engineering, Faculty of Computer Science and Engineering, Hodeidah University, Hodeidah, Yemen

Received: 28.04.2019 • **Accepted/Published Online:** 12.10.2020 • **Final Version:** 27.01.2021

Abstract: The potential capability to automatically detect and classify human behavior as either normal or abnormal events is an important aspect in intelligent monitoring/surveillance systems. This study presents a new high-performance framework for detecting behavioral abnormalities in video streams by utilizing only the patterns for normal behaviors. In this paper, we used a hybrid descriptor, called a foreground optical flow energy (FGOFE), which makes use of two effective motion techniques in order to extract the most descriptive spatiotemporal features in video sequences. The FGOFE descriptor can effectively capture both weak and sudden incidents in a scene. The sequential generalization of *k*-means (SGK) algorithm was applied in this study to generate the dictionary set that can sparsely represent each signal; in addition, the orthogonal matching pursuit algorithm was utilized to recover high-dimensional sparse features when referring to a few numbers of noisy linear measurements. Using the SGK allows gaining a less complex and quicker implementation compared to other dictionary learning methods. We conducted comprehensive experiments to analyze and evaluate the ability of our framework in detecting abnormalities using several public benchmarks, which contain different abnormal samples and various contextual compositions. The experimental results show that the proposed framework achieved high detection accuracy (up to 95.33%) and low frame processing time (31 ms on average) compared to the relevant related work.

Key words: Abnormal detection, video surveillance, sparse representation, sequential generalization of *k*-means, principal component analysis, orthogonal matching pursuit

1. Introduction

Due to the increasing demand for public security measures [1, 2], abnormal behavior detection in data streams is a rich area of active research in computer vision and machine learning communities. The term "abnormal behavior" refers to a suspicious action which can potentially threaten human life, health, and public safety. Abnormal behavior is neither a specific activity (e.g., hands raised up, sitting down, standing up), nor a simple behavior (e.g., jumping, running, cycling). It is a complex behavior, which may include several simple behaviors and a series of actions [3, 4].

Abnormal behavior detection is a challenging process due to many factors, including light conditions, occlusions between individuals, quality of the video, camera motion, complexity of backgrounds, small size of the abnormal incident, lack of a sufficient amount of abnormal ground truth training data, and the density of

*Correspondence: kmaahlam2@live.utm.my

crowds [5]. A human being may be monitored as noncrowds (individuals, a group of people) or crowds. Most of the existing studies for abnormal behavior detection based on individuals have been used for guaranteeing the safety of old and impaired people in medical centers, nursing houses, or infirmaries. In addition, a large body of studies concentrates on identifying abnormalities in behavior regarding breaking the law or violation of security or incidents relating to safety issues. Other studies focus on the detection of infrequent incidents and can be applied to undefined applications. Abnormal behavior detection and analysis in crowded environments is also a significant topic. All the behaviors that deviate from the surrounding pedestrian activities are usually deemed to be abnormal incidents [1].

Sparse representations are effectively applied in different significant applications related to signal and image processing. These include image denoising, coding, echo channels modeling, activity recognition, and abnormal detection [6]. The overcomplete bases utilized for signal representation are either constant (e.g., by obtaining rows of widespread transformations such as discrete cosine or wavelets) or learned (by applying a representative pattern of the signals). Our study deals with the latter (which is also referred to as dictionary learning). In this paper, we put forward a new high-performance abnormal behavior detection framework suitable for both crowd and noncrowd video streams. The main aspects of our framework comprise: (*i*) proposal of a new and effective framework for abnormality detection. This can generate the best dictionary set using sequential generalization of k-means (SGK) algorithm, which provides a less complex and faster implementation compared to other dictionary learning methods. To the best of our knowledge, this is the first work that addresses abnormal behavior detection problem using SGK, (*ii*) proposal of a novel descriptor called foreground optical flow energy (FGOFE) to detect abnormal behaviors in surveillance videos. This is performed by making use of two effective techniques, namely background subtraction and optical flow energy, so as to extract highly descriptive spatiotemporal characteristics. We have conducted comprehensive experiments to analyze and evaluate the ability of our framework in detecting abnormalities. The process uses three public benchmarks which contain different abnormality incidents and contextual compositions.

The rest of this paper is organized as follows. Section 2 reviews the related work, while Section 3 elaborates on the proposed framework in detail. In Section 4, the experimental results, along with performance analysis of the proposed framework, are discussed. Finally, the paper is concluded in Section 5.

2. Related work

Activity recognition models have recently been utilized in a number of multidisciplinary research studies, focusing especially on abnormal behavior detection. This section will elaborate on two significant aspects in abnormal detection area, specifically: feature extraction and sample modeling approaches. For feature representation, approaches of abnormal detection can be classified into high-level feature-based and low-level feature-based [1, 7]. In the first category, identification or object tracking is carried out in order to utilize an object's trajectory to detect abnormal incidents [8, 9]. The approach in [8] uses an adaptive background subtraction to produce an outdoor real-time tracker to suit trajectories of objects. In [9], a tracking-based method of the kernel-based object is used to show moving objects.

Although high-level feature-based approaches can plainly represent an object's spatial status by the time, this status is influenced by tracking errors, occlusions, and noises. This leads, in turn, to failure in the abnormal detection task in both cluttered and crowded scenes. To avoid this limitation, recent studies seek at the pixel level to extract the low-level motion features from the frames. In [10], motion features were represented by clustering space-time interest points to generate a bag-of-visual word (BoW) model. Lucas and Kanade (LK)

technique was used in [5, 11] to calculate optical flow for each pixel. To create a bag of words depicting videos, every incident is appointed by the position and direction of motion of a nonoverlapping unit.

In terms of sample modeling, abnormal detection approaches can be classified into the following types: dynamic Bayesian network (DBN) [12, 13], Bayesian topic models (BTMs) [14], clustering-based models [8, 15], artificial neural network (ANN) [16, 17], deep learning-based models [18], and sparse representation-based models [19–21]. The most commonly used among those types are the DBN and BTMs, which can dynamically model human behavior. The hidden Markov model (HMM) is the common type of DBN models [12]. In [13], to model the extracted trajectories, a continuous k-state HMM was utilized. To avert the high cost of DBN computation, BTMs were proposed [7]. In [14], a probabilistic latent semantic analysis model was put forward to determine the latent topics (in that if there are no topics that demonstrate the detected words of a clip, that video clip is considered to be abnormal). Classification based on a tree-structured approach in [8] was used to cluster a codebook into multiple activity samples. Authors in [15] employed a multivariate Gaussian mixture model (MGMM) to model objects' respective size and speed.

ANN approaches build neurons in order to model behavior samples. Authors in [16] utilized the interaction of energy potentials to discover prominent points with neighboring objects. For modeling the behavioral samples in crowd sites, authors in [17] proposed a self-organizing online map created by a 2-D lattice. Observed behavior is detected as an anomaly if its winning neuron distance is greater than a certain threshold. Recently, deep learning models have been employed for detecting anomalies [18]. In these models, there are no obvious processes for extracting features as well as the representations learned by the network itself. Nevertheless, they evoke a comparatively huge amount of data for the training process to avoid overfitting. Moreover, these models are computationally expensive in terms of time and resources [7]. In fact, real-world surveillance video streams have several forms of abnormal incidents, which are hard to define and annotate. In other words, the strength of deep learning is more prominent in case the training dataset is very large. However, this advantage may not be much efficient in some particular domains such as abnormal behavior detection in surveillance videos, which may not have an adequately large amount of data. Apart from deep learning models, the proposed framework in this study can achieve reliable and timely performance without counting on large-size datasets, long processing time, and a great deal of annotation [22, 23].

In recent years, sparse representation-based approaches have attracted considerable attention [1, 7]. Cong et al. [19] proposed a new descriptor named multiscale histogram of optical flow and used sparse reconstruction costs over the normal bases to build an abnormal detection model. A new trajectory-based method was proposed in [20], which classifies motion events using sparse representation classification. Inspired by [21], the proposed framework made use of dictionary learning to detect abnormalities.

3. The proposed framework

In this paper, we propose a new abnormal behavior detection framework based on salient spatiotemporal features (which are extracted using the hybrid FGOFE descriptor), as well as SGK and orthogonal matching pursuit (OMP) techniques. The block diagram of the proposed framework for surveillance videos is illustrated in Figure 1. The framework is composed of two main components, namely, a training phase and a testing phase. Each of these consists of three subcomponents. More specifically, the training stage consists of the three following subcomponents: (*i*) behavior feature extraction, background subtraction features based on Gaussian mixture model algorithm, and optical flow energy features are calculated in each volume, (*ii*) principal

component analysis (PCA) algorithm for dimensionality reduction, (*iii*) sequential generalization of *k*-means (SGK) algorithm to generate the best dictionary set that can sparsely represent each feature. Likewise, the testing stage consists of three subcomponents: (*i*) feature extraction as in the training stage, (*ii*) the orthogonal matching pursuit (OMP) algorithm to reconstruct high-dimensional sparse features regarding the few numbers of noisy linear measurements, and (*iii*) reconstruction error (RE) detection to decide whether the incident is abnormal by using a certain threshold.

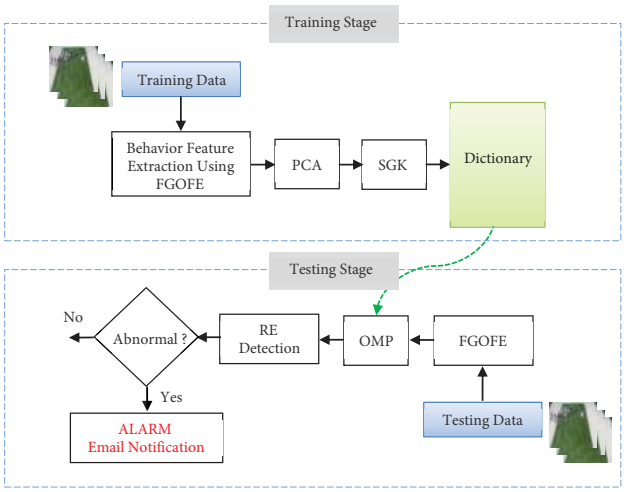


Figure 1. Block diagram of the proposed framework.

3.1. Foreground optical flow energy (FGOFE)

In this section, the process of feature extraction using FGOFE for abnormal detection is explained. In this study, we utilized motion algorithms for extracting a set of descriptive features. In particular, inspired by [24], we employed two effective motion algorithms, specifically, background subtraction and optical flow.

Figure 2 illustrates the feature extraction process using FGOFE. Given an input video, all video frames were resized to 160×120 pixels. Each frame was then divided into a set of nonoverlapping 10×10 blocks, with each region containing a spatiotemporal block for five consecutive frames. Following this, the salient spatiotemporal features were extracted using background subtraction and optical flow energy on the spatiotemporal blocks. The salient spatiotemporal features were then normalized such that the expected mean and variance fall within 0 and 1. Those salient spatiotemporal features represent the motion information.

After applying the background subtraction and optical flow energy, the extracted features were filtered to remove unwanted small and noisy objects. Thus, two feature sets B_T and O_E were constructed using background subtraction and optical flow energy, respectively. These features were then concatenated into one feature set, namely : $B_T O_E$.

3.2. Foreground features

To obtain foreground features, any background subtraction algorithm (BS) similar to those in [25, 26] can be used provided that it is not computationally expensive, in terms of time. BS can quickly identify regions of interest that can serve as masks for more evolved algorithms [27]. To detect moving objects in real-time, a

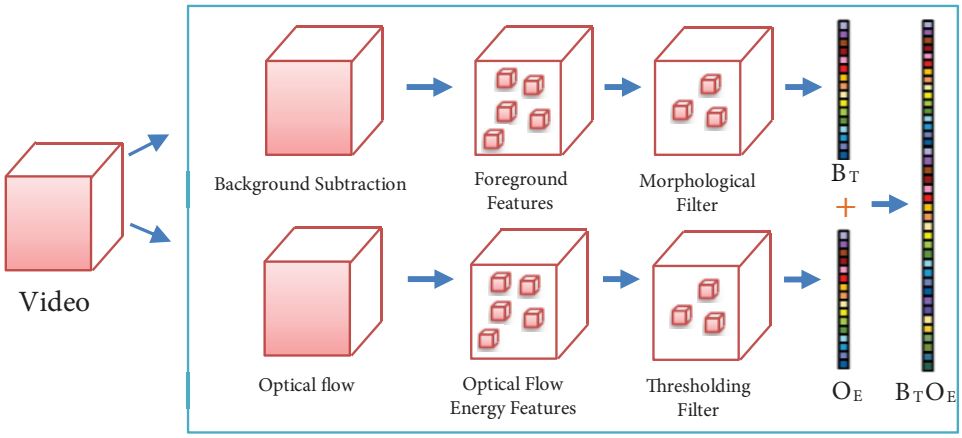


Figure 2. Feature extraction using FGOFE.

single-Gaussian background updating method (SGBUM) was utilized. SGBUM is considered an alternative to the more intricate mixture Gaussian model for extracting foreground features at a high processing rate. For each frame (where (i, j) is the pixel location) for the time duration t in a video frame in the SGBUM, the gray level mean and variance, μ_T and σ_T, respectively, represent the parameters. These two statistical values can be distinctly computed by removing the last frame in the historical frame series and adding the present frame for continuously monitoring. For N consecutive frames $\{f_t, t = T, T-1, ..., T-(N+1)\}$, where T is the present frame time, the mean and variance of the SGBUM at time frame T were calculated as per the following equations:

$$\mu_T = E[f(i,j)] = \frac{1}{N} C_T(i,j) \tag{1}$$

$$\sigma_T^2 = E[f^2(i,j)] - \{E[f(i,j)]\}^2 = \frac{1}{N} C_T^2(i,j) - \mu_T^2 \tag{2}$$

where:

$$C_T(i,j) = \sum_{k=0}^{N-1} f_{T-k}(i,j) = C_{T-1}(i,j) - f_{T-N}(i,j) + f_T(i,j) \tag{3}$$

$$C_T^2(i,j) = \sum_{k=0}^{N-1} f_{T-k}^2(i,j) = C_{T-1}^2(i,j) - f_{T-N}^2(i,j) + f_T^2(i,j) \tag{4}$$

By dropping the last frame $(f_{T-N}(i,j))$ as well as adding the present frame $(f_T(i,j))$ to the frame series, $C(i,j)$ and $C_{T(}^2(i,j)$ can be updated effectively. Thus, the updating calculation encompasses only two basic operations. Accordingly, a very high frame processing rate can be accomplished. It is worth mentioning that the processes of updating the mean and variance are invariant to the number of frames. Note that for motion detection in video streams, the frames of background will present a roughly identical gray level with a small variance value. On the other hand, the gray value of the foreground pixels will be clearly dissimilar from the background pixels. The lower and upper control limits to detect foreground pixels in the current frame $f_T(i,j)$ can be obtained by $\mu_{T-1}(i,j) \pm u.\sigma_{T-1}(i,j)$, in which u is considered as a control constant. In

case the gray level of $f_T(i,j)$ is out of these limits, the pixel is then deemed to be a foreground feature. The result of the detection process is represented by a binary mask $M_T(i,j)$, and calculated according to equation (5).

$$M_T(i,j) = \begin{cases} 0 & if\ |f_T(i,j) - \mu_{T-1}(i,j)| \leq u \cdot \sigma_{T-1}(i,j) \\ 1 & Otherwise \end{cases} \tag{5}$$

In this study, because the gray values between the background point (0) and the foreground point (1) are usually clearly dissimilar, the control constant u is limited to 5. Figure 3 shows an example of the extracted foreground features using the proposed framework.

Figure 3. An example frame of foreground features on UCSD PED1 dataset. The picture on the left is an original frame, and the picture on the right is the binary mask of foreground features.

3.2.1. Optical flow energy features

In some scenes, incidents like panic and sudden actions may not be correctly described by their long-term activity, the number of objects, or even their size. Instead, they may be defined by their motion's speed. Therefore, to capture the sudden actions in a scene, this study employed the optical flow energy process. For each space location (i_p, j_p) at the frame difference t_p, the optical flow energy is calculated according to equation (6) below:

$$O_E(i_p, j_p, t_p) = \frac{1}{N} \sum_{n=1}^{N} \left\| v_i^{(n)}, v_j^{(n)} \right\|_2 \tag{6}$$

such that, for N pixels in a video volume, v_i and v_j are the horizontal and vertical components of optical flow [24].

3.3. Principal component analysis (PCA)

Principal component analysis was adopted in the proposed framework for projecting a high number of extracted features into a lower-dimensional space. As shown in Figure 4a, PCA uses input data, for example, A to obtain new values $B_{ij}(i = 1, 2, ..., m$ and $j = 1, 2, ..., k)$ that are combinations of the inputs. The new

values are uncorrelated and ordered so that B_{11} has the greatest variance and B_{mk} the lowest. By preserving only B_{ij} values that have the greatest variance, the original features can be represented in lower dimensions. In this study, we compute PCA using singular value decomposition (SVD). SVD aims to diagonalize the data matrix $A \in R^{p \times q}$ into three matrices [28] as described below:

$$A_{n \times m} = U_{n \times n} S_{n \times m} V^T_{m \times m} \tag{7}$$

where $U_{n \times n}$ represents the left singular vectors, $S_{n \times m}$ is a diagonal matrix that represents singular values that are sorted descendingly, and $V_{m \times m}$ denotes the right singular vectors. The left and right matrices (i.e. U and V) are orthonormal bases. To compute SVD, firstly, S and V^T are computed by diagonalizing $A^T A$ as in the following: $A^T A = (USV^T)^T (USV^T) = US^2 V^T$, where $U^T U = I$. U is then computed as follows: $U = AVS^{-1}$. The columns of V matrix represent the eigenvectors of $A^T A$, which are the principal components of the PCA as shown in Figures 4b and 4c. It is worth mentioning that, because the number of principal components and their eigenvalues equalize q, the dimension of the original data must be reversed to be compatible with the SVD manner. In other words, the mean-centering matrix is transposed before computing the SVD; thus, each sample is represented by one row as shown in Figure 4a.

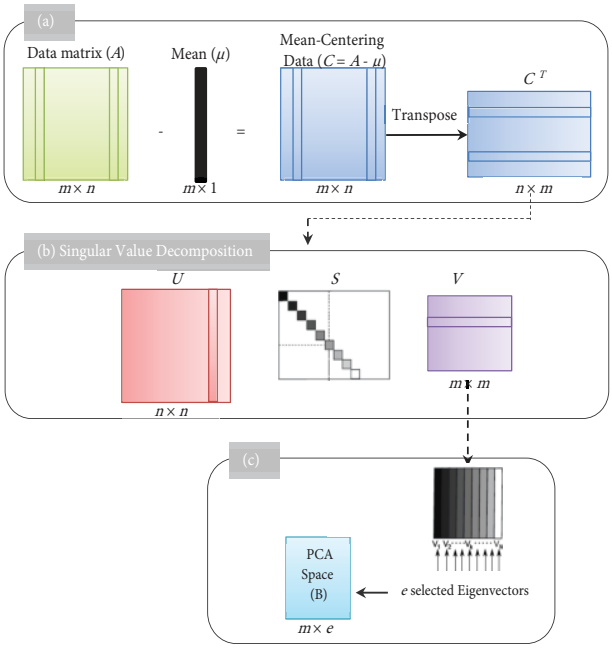


Figure 4. Dimensionality reduction using PCA via SVD.

3.4. Dictionary learning with SGK

Lately, dictionary learning algorithms for sparse representation are very similar to those of k-means clustering. The k-means algorithm [29, 30] aims to find k clusters, in which each cluster is illustrated precisely by one of the k centroids. These k points create a codebook matrix $C = [c_1, c_2, ..., c_k]$. Given a set of signals $Y = [y_1, y_2, ..., y_n]$ and C, the main target of k-means algorithm is to put each signal y_i to the nearest centroid, with the help of the minimum l_2 norm distance. Consequently, a vector α_j is generated. This

is a vector containing zeros except at the position jth whose assigned value is 1. Equation (8) below is used to build such vector:

$$min_{C,X}\{\|Y - CX\|_F^2 \; s.t. \forall i, \; x_i = \alpha_k \text{ for some } k\} \tag{8}$$

The coefficient vector α_j for a signal y_i has only one '1' and the remaining entries are '0s'. Expanding upon this, if more than nonzero coefficients are used, and values other than 1 are permitted, this turns to a more general algorithm called K-singular value decomposition (K-SVD). In K-SVD and other learning dictionary learning algorithms, the codebook is entitled dictionary *(D)* and its columns are called atoms. Nowadays, K-SVD has become well-known and the most commonly used dictionary learning algorithm [31].

The performance of SGK's training is comparable to those of K-SVD. Furthermore, it has achieved faster and more efficient implementation [32]. The method of optimal directions (MOD) decreases $\|Y - DX\|$ in a direct way, which is considered a convex issue for a presented X. The K-SVD adjusts the atoms sequentially by applying singular value decomposition (SVD) as well as altering the representations with the atoms. The SGK algorithm optimizes the atoms sequentially, without changing the representations, by solving a least-squares problem. The aim of SGK is to decrease the sparse representation through sequentially updating the atoms $d_k \in R^n$ for $k = 1, ..., K$. As in [33], SGK is incorporated effectively via sparse representation into the framework of image denoising. In this paper, SGK is dedicated to work for the purpose of abnormal detection in video streams.

After applying PCA on the extracted features, the dictionary should be trained by the statements of these features. The dictionary updating is demonstrated as follows:

$$arg\, min_D(\|f(y_i - DX_i)\|_2 + \lambda \|X_i\|_1) \tag{9}$$

It is worth noting that SGK can find not only the least square solution but also the sparse representation [34]. Thus, in this paper, we use SGK to train the dictionary. Like dictionary training algorithms, the error matrix E_k corresponding to d_k for extracted features $F(Y)$ should be computed in SGK.

$$E_k = F(Y) - \sum_{j \neq k} d_j a_j \tag{10}$$

where a_j is the j^{th} row of X. By substituting (10) in (9), the Lagrange function can be obtained as follows:

$$minimize\, L_k = \|E - d_k a_k\|_F + \sum_{i=1}^{N} \lambda \|X_i\|_1 \tag{11}$$

3.5. Sparse signal reconstruction using orthogonal matching pursuit (OMP)

OMP is an iterative greedy approach [6], which picks out at each step the column that is most correlated with the current residuals. It aims to orthogonally project the observation onto the linear subspace spanned by the columns, which have already been chosen. OMP updates the residuals and then iterates. Compared with other alternative greedy approaches (such as matching pursuit (MP) and stagewise orthogonal matching pursuit (StOMP)), OMP is simple and computationally inexpensive in terms of time [35]. A detailed algorithm of OMP is described in Algorithm 1. OMP behaves greedily when selecting the atoms. That is, OMP selects an atom

that is linearly independent of the atoms already selected as shown in equation (12) below:

$$\langle r_l, d_{hj} \rangle = 0 \text{ for } j = 1, ..., l \tag{12}$$

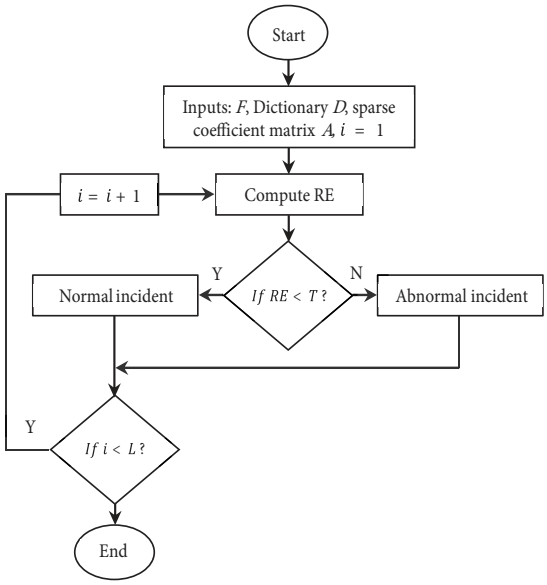
Consequently, in N steps, the residual must be equal to zero. Moreover, the atoms are identical to the index set in producing a full rank matrix; therefore, the solution of least squares is unique. Because the residual is orthogonal to the atoms selected, OMP guarantees that each atom is selected only once in the approximation. The least squares solution obtains the approximation by:

$$y_l = \sum_{j=1}^{l} A[h_i] d_{hl} \tag{13}$$

3.6. Abnormal detection testing based on SGK's dictionary and OMP coefficients

Figure 5 represents the flowchart of the testing stage. In the training stage, we generated the dictionary $D = \{d_1, ..., d_L\}$. For new testing features, F, in the testing stage, the sparse coefficient matrix, $A = \{a_1, ..., a_L\}$, was calculated using OMP. Following this, D was then examined to determine whether it has a combination to render its reconstruction error (RE) lower than a certain threshold T. If D has such combination, it is considered a normal incident; otherwise, it is deemed to be an abnormal incident. It can be simply achieved by checking the RE of each column in D, as shown in the following equation:

$$RE = min \left\| F - d_i a_i \right\|_2^2 \; \forall i = 1, ..., L \tag{14}$$



Algorithm 1 OMP algorithm.

Input: Testing features $y \in \mathbb{R}^{\mathbb{N}}$, Dictionary: D (its columns MUST be normalized), the max number of coefficients for each signal L (*set it to 10*).
Output: Sparse coefficient matrix $A \in R^h$.
Initialize: initial residual $r_0 = y$
index set $in = \{\}$
Loop index $l = 1$
while convergence not reached

- Find an index $h_1 : h_1 = argmax_k \left| \langle r_{(l-1)}, d_h \rangle \right|$
- Update the index set $in \leftarrow in \cup h_1$
- Calculate $A = argmin_A \left\| y - \sum_{j=1}^{l} A[h_j] d_{hj} \right\|_2$
- Calculate the new residual $r_l = y - \sum_{j=1}^{l} A[h_j] d_{hj}$
- Update the loop counter $l = l + 1$

Figure 5. Flowchart of the testing stage.

4. Experimental results and discussion

Experiments were carried out using three common dataset benchmarks to prove the ability of the proposed framework for detecting behavior abnormalities. In the next subsections, the benchmarks are first introduced

and then the evaluation criteria are explained. Finally, the experimental results and comprehensive evaluation are presented.

4.1. Parameters setting

In the experiments, the parameters that need to be set comprise the PCA compression dimension $PCAdim$, dictionary size K, the max number of coefficients for each signal L to use OMP algorithm, and the learning rate η to extract foreground features. In our implementation, we set the parameters as $PCAdim = 100$, $K = 94$, $L = 10$, and $\eta = 10^{-2}$, where these values are empirically determined. Undoubtedly, the results of the experiments are sensitive to these mentioned parameters.

4.2. Dataset benchmarks for abnormal detection

In order to evaluate the performance of the proposed framework for abnormal behavior recognition, it was tested on both crowded and uncrowded views, which contain one or two individuals in the view. Three publicly available datasets of abnormal samples were utilized, specifically: (*i*) Avenue Dataset [21]: This dataset has 16 and 21 video clips for training and testing, respectively. The total number of frames is 30,652. There are fourteen irregular events comprising loitering, running, throwing objects, and walking in the opposite direction. Figures 6a–6c illustrate some incidents from this dataset; (*ii*) UCSD Dataset [36]: This dataset consists of two scenes, PED1 and PED2. Each scene has several training and testing video clips, in which training sets have only normal samples and testing sets have both normal and abnormal samples. It is worth mentioning that PED1 is more challenging than PED2 because the camera angle produces a larger perspective distortion. Moreover, the abnormal samples in PED1 involve not only abnormalities perpetrated by small carts, bikers, and skateboarders etc., but also contextual abnormalities such as a person walking over the grass. Figure 6d–6f represent some frames of normal and abnormal patterns from the UCSD-PED1 dataset; and (*iii*) UMN Dataset [37]: This is a commonly used benchmark. It comprises eleven video footages for three different escape views (one indoor view and two outdoor views). The total length for this dataset is 7741 frames. In addition, the resolution of the frames is 320×240 pixels. Figures 6g and 6h show some frames of normal and abnormal patterns from the UMN-Indoor scene dataset.

4.3. Experimental setup and evaluation criteria

The experiments in this study were conducted using MATLAB R2017b (9.3.0.713579) x64 on Linux platform with an Intel Core i7-4600U CPU working at 2.10 GHz with a 4 MB cache and 8 GB RAM. In addition, we used OpenCV Matlab library executed by Alalek to extract the foreground features[1]. This gives us faster implementation other than built-in MATLAB function. Any abnormal detection framework can be evaluated at the frame-level; a frame-level–based criterion was adopted here to indicate that a frame can be deemed suspicious if it consists of any anomaly pixel, irrespective of its location. The frame-level–based receiver operating characteristic (ROC) curve and area under the curve (AUC) were utilized in this study as the performance evaluation criterion. Furthermore, we used the confusion matrix to measure detection accuracy, recall, and Fscore values in the experimental datasets. The structure of the confusion matrix for classification of abnormal detection algorithms is represented in Table 1 as follows: true-negative (TN): The number of normal samples that are correctly detected as normal; false-negative (FN): The number of abnormal samples incorrectly detected as normal; true-positive (TP): The number of abnormal samples that are correctly detected as abnormal; and

[1]The code provided at https://github.com/opencv/opencv_contrib/tree/master/modules/matlab

Figure 6. (a, b, c): Some patterns from Avenue dataset: (a) Normal walking, (b) Abnormal behavior of running, (c) Abnormal behavior of throwing objects. (d, e, f): Some patterns from UCSD-PED1 dataset: (d) Normal walking behavior, (e) Abnormal behavior of biking, (f) Abnormal behavior of skating. (g, h): Some patterns from UMN-Indoor scene dataset: (g) Normal walking, (h) Escape in panic behavior.

false-positive (FP): The number of normal samples that are incorrectly detected as abnormal. The computation formulas of accuracy (Acc), recall, Fscore, and equal error rate (EER) measures are provided as in the following formulas [38]:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \qquad Recall = \frac{TP}{TP + FN}$$

$$Fscore = \frac{2TP}{2TP + FP + FN}, \qquad EER = \frac{FP + FN}{TP + TN + FP + FN}$$

Table 1. Confusion matrix for classification.

Actual	Detected	
	Normal	Abnormal
Normal	TN	FP
Abnormal	FN	TP

4.4. Results and comparisons with state-of-the-art methods

In this paper, we propose an abnormal behavior detection framework in surveillance videos based on spatiotemporal features, SGK, and OMP. Figure 7 shows examples of abnormal frames detected by our framework, where Avenue dataset (in the first row) represents the following abnormal incidents: a child jumping, a man running and throwing objects, respectively; UMN dataset (in the second row) shows people running in panic in three scenes (Lawn, Indoor, and Plaza); UCSD PED1 (in the third row) demonstrates cyclists, and small cars, respectively. We used FGOFE to extract visual features and SGK was then used to generate the abnormal detection model. Our framework improved the abnormal detection method in [21], based on sparse combination algorithm, which uses a spatiotemporal gradient model to extract the salient spatiotemporal features. We compared our framework with [21] in terms of accuracy, frame processing time (FPT), recall, and Fscore. The experimental results employing avenue dataset are shown in Table 2. It can be seen that the average accuracy of our framework has been improved, which indicates that our framework can detect more abnormal incidents than [21]. Furthermore, as shown in Figures 8a and 8b, the recall and Fscore values were significantly increased in most video clips of avenue dataset. The recall value in some clips reached up to 98%. Table 3 represents the experimental results conducted on UMN dataset in terms of the AUC, accuracy, and EER, respectively.

Table 2. The overall accuracy, frame processing time, recall, Fscore experimental results between Lu et al.'s method and ours on avenue dataset.

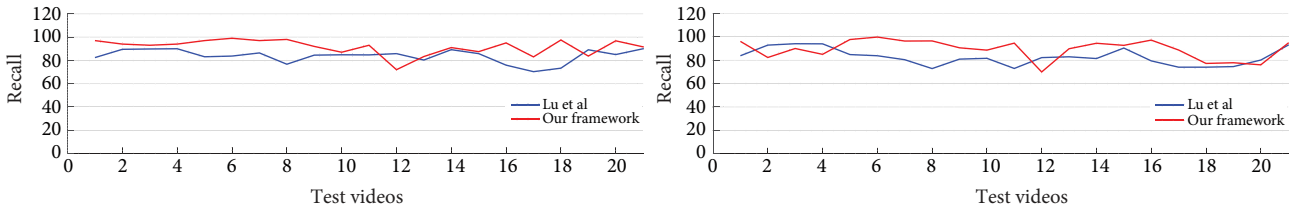
Method	Accuracy (%)	FPT (ms)	Recall (%)	Fscore (%)
Lu et al. [21]	76.6560	6	83.8	83.8
Proposed	96.5586	13	91.0	89.0

Table 3. Experimental results of the proposed framework using UMN dataset.

Dataset	AUC	Accuracy (%)	EER
UMN (Lawn scene)	0.9449	93.9100	0.0609
UMN (Indoor scene)	0.9211	95.3300	0.0420
UMN (Plaza scene)	0.9429	93.6667	0.0391
Avg.	0.9363	94.3022	0.0473

A quantitative comparison of different methods for abnormal detection at frame-level using UMN dataset is shown in Figures 9a and 9b. As can be clearly in Figure 9, our framework demonstrates better performance results simultaneously in terms of the balancing of both the AUC and the frame processing time values than [19, 21, 22, 39–42].

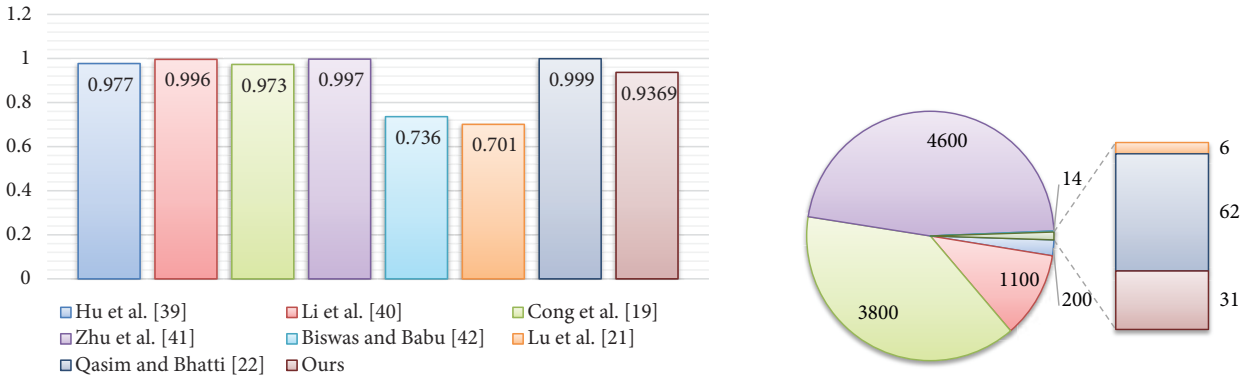
Figure 7. Examples of abnormal frames detected by our framework.



(a) Recall measure (b) Fscore measure

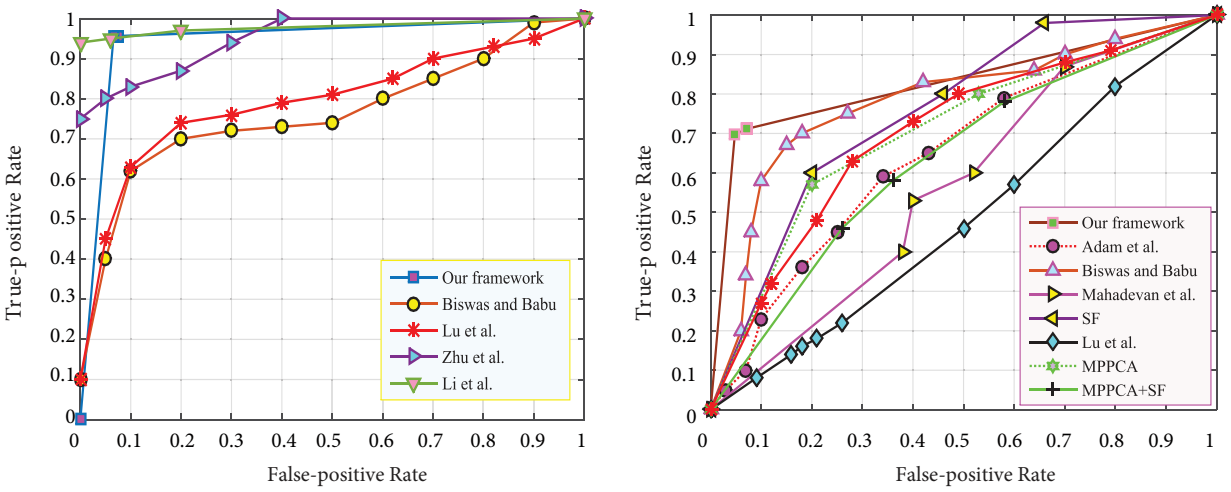
Figure 8. Comparison of Lu et al.'s method and our proposed framework using Avenue dataset.

The frame-level ROC curves applying UMN and UCSD-PED 1 are represented in Figures 10a and 10b, respectively. When the false positive value is small, our framework has a relatively high detection rate, and it can be absolutely vital for developing realistic abnormal detection systems. It can be observed that the proposed framework performs better than [21, 36, 37, 42–44], and provides competitive results compared to [40, 41].



(a) AUC (b) Processing time (ms)

Figure 9. A comparison of different methods for abnormal detection at frame level using UMN dataset.



(a) Using UMN dataset (b) Using UCSD-PED1 dataset

Figure 10. ROC curve for the proposed framework with the state-of-the-art algorithms. Abbreviation of the methods: Biswas and Babu [42], Lu et al. [21], Zhu et al. [41], Li et al. [40], Adam et al. [43], Mahadevan et al. [36], SF [37], MPPCA [44], MPPCA-SF [36].

5. Conclusion

Generally, the detection of abnormal human behavior in video streams (in particular in surveillance videos), can be considered to be one of the most important subjects that attract growing attention from researchers in the field of maintaining public safety. So far, the existing results obtained from the previous related work have encouraged more studies to improve the overall performance without degrading computation time and complexity.

We have used a hybrid descriptor in this paper, called a foreground optical flow energy (FGOFE), using two effective motion techniques to extract the most descriptive spatiotemporal features in video sequences. The sparse representation has evolved and has been applied broadly to fields of machine learning and computer vision. Therefore, in this study, a new effective framework to detect abnormalities based on SGK and OMP algorithms has been presented. The SGK algorithm was employed to generate the best dictionary set that

can sparsely represent each feature, while the OMP was used to recover high-dimensional sparse features when referring to a few numbers of noisy linear measurements. The advantages of using SGK's learning lie in the provision of a less complex and quicker implementation compared with other dictionary learning methods. The experimental evaluation represents our framework as achieving a very competitive performance for the state-of-the-art methods in all mentioned performance criteria such as detection accuracy, EER, and frame processing time. Future work of this study is to improve the detection accuracy of our framework further and maintain the frame processing time as little as possible to achieve online performance. Also, results achieved so far strongly urge future work for the sake of further studies with other real-world datasets to test and improve the proposed framework.

References

[1] Al-Dhamari A, Sudirman R, Mahmood NH. Abnormal behavior detection in automated surveillance videos: a review. Journal of Theoretical & Applied Information Technology 2017; 95(19): 5245-5263.

[2] Sahu AK, Sahu M. Digital image steganography techniques in spatial domain: a study. International Journal of Pharmacy & Technology 2016; 8(4): 5205-5217.

[3] Mu C, Xie J, Yan W, Liu T, Li P. A fast recognition algorithm for suspicious behavior in high definition videos. Multimedia Systems 2016; 22(3): 275-285.

[4] Al-Dhamari A, Sudirman R, Mahmood NH, Khamis NH, Yahya A. Online video-based abnormal detection using highly motion techniques and statistical measures. Telkomnika 2019; 17(4): 2039-2047.

[5] Gnouma M, Ejbali R, Zaied M. Abnormal events detection in crowded scenes. Multimedia Tools and Applications 2018; 1: 1-22.

[6] Sadeghi M, Babaie-Zadeh M, Jutten C. Learning overcomplete dictionaries based on atom-by-atom updating. IEEE Transactions on Signal Processing 2014; 62(4): 883-891.

[7] Yu B, Liu Y, Sun Q. A content-adaptively sparse reconstruction method for abnormal events detection with low-rank property. IEEE Transactions on Systems, Man, and Cybernetics: Systems 2017; 47(4): 704-716.

[8] Stauffer C, Grimson WEL. Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000; 22(8): 747-757.

[9] Bruni V, Vitulano D. An improvement of kernel-based object tracking based on human perception. IEEE Transactions on Systems, Man, and Cybernetics: Systems 2014; 44(11): 1474-1485.

[10] Wu Q, Wang Z, Deng F, Chi Z, Feng DD. Realistic human action recognition with multimodal feature selection and fusion. IEEE Transactions on Systems, Man, and Cybernetics: Systems 2013; 43(4): 875-885.

[11] Fu W, Wang J, Lu H, Ma S. Dynamic scene understanding by improved sparse topical coding. Pattern Recognition 2013; 46(7): 1841-1850.

[12] Felori F, Margez F. Motif mining by combined hidden markov model and clustering method. Journal of Bioinformatics and Intelligent Control 2015; 4(1): 35-43.

[13] Porikli F, Haga T. Event detection by eigenvector decomposition using object and frame features. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop; Washington, DC, USA; 2004. pp. 114-114.

[14] Varadarajan J, Odobez JM. Topic models for scene analysis and abnormality detection. In: 12th IEEE International Conference on Computer Vision Workshops (ICCV); Kyoto, Japan; 2009. pp. 1338-1345.

[15] Basharat A, Gritai A, Shah M. Learning object motion patterns for anomaly detection and improved object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Anchorage, AK, USA; 2008. pp. 1-8.

[16] Cui X, Liu Q, Gao M, Metaxas DN. Abnormal detection using interaction energy potentials. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Colorado Springs, CO, USA; 2011. pp. 3161-3167.

[17] Feng J, Zhang C, Hao P. Online learning with self-organizing maps for anomaly detection in crowd scenes. In: 20th International Conference on Pattern Recognition (ICPR); Istanbul, Turkey; 2010. pp. 3599-3602.

[18] Shao J, Loy CC, Kang K, Wang X. Crowded scene understanding by deeply learned volumetric slices. IEEE Transactions on Circuits and Systems for Video Technology 2017; 27(3): 613-623.

[19] Cong Y, Yuan J, Liu J. Sparse reconstruction cost for abnormal event detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Colorado Springs, CO, USA; 2011. pp. 3449-3456.

[20] Li C, Han Z, Ye Q, Jiao J. Abnormal behavior detection via sparse reconstruction analysis of trajectory. In: Sixth International Conference on Image and Graphics (ICIG); Hefei, Anhui, China; 2011. pp. 807-810.

[21] Lu C, Shi J, Jia J. Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE International Conference on Computer Vision; Sydney, NSW, Australia; 2013. pp. 2720-2727.

[22] Qasim T, Bhatti N. A low dimensional descriptor for detection of anomalies in crowd videos. Mathematics and Computers in Simulation 2019; 166: 245-252.

[23] Srinivasan A, Gnanavel VK. Multiple feature set with feature selection for anomaly search in videos using hybrid classification. Multimedia Tools and Applications 2019; 78(6): 7713-7725.

[24] Leyva R, Sanchez V, Li CT. Video anomaly detection with compact feature sets for online performance. IEEE Transactions on Image Processing 2017; 26(7): 3463-3478.

[25] KaewTraKulPong P, Bowden R. An improved adaptive background mixture model for real-time tracking with shadow detection. In: Remagnino P, Jones GA, Paragios N, Regazzoni CS (editors). Video-Based Surveillance Systems. Boston, MA: Springer, 2002, pp. 135-144.

[26] Elgammal A, Duraiswami R, Harwood D, Davis LS. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proceedings of the IEEE 2002; 90(7): 1151-1163.

[27] Abdallah ACB, Gouiffès M, Lacassagne L. A modular system for global and local abnormal event detection and categorization in videos. Machine Vision and Applications 2016; 27(4): 463-481.

[28] Tharwat A. Principal component analysis-a tutorial. International Journal of Applied Pattern Recognition 2016; 3(3): 197-240.

[29] Kutbay U, Ural AB, Hardalaç F. Underground electrical profile clustering using K-MEANS algorithm. In: 23nd Signal Processing and Communications Applications Conference (SIU); Malatya, Turkey; 2015. pp. 561-564.

[30] Hardalaç F, Kutbay U, Şahin İsa, Akyel A. A novel method for robust object tracking with k-means clustering using histogram back-projection technique. Multimedia Tools and Applications 2018; 77(18): 24059-24072.

[31] Zhou T, Liu F, Bhaskar H, Yang J, Zhang H et al. Online discriminative dictionary learning for robust object tracking. Neurocomputing 2018; 275: 1801-1812.

[32] Sahoo SK, Makur A. Sparse sequential generalization of k-means for dictionary training on noisy signals. Signal Processing 2016; 129: 62-66.

[33] Sahoo SK, Makur A. Image denoising via sparse representations over sequential generalization of k-means (SGK). In: 9th International Conference on Information, Communications and Signal Processing (ICICS); Tainan, Taiwan; 2013. pp. 1-5.

[34] Lu G, Zhang K, Huang S, Zhang Y, Feng Z. Modulation recognition for incomplete signals through dictionary learning. In: IEEE Wireless Communications and Networking Conference (WCNC); San Francisco, CA, USA; 2017. pp. 1-6.

[35] Ren H, Pan H, Olsen SI, Moeslund TB. A comprehensive study of sparse codes on abnormality detection. arXiv preprint arXiv:1603.04026 2016.

[36] Mahadevan V, Li W, Bhalodia V, Vasconcelos N. Anomaly detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); San Francisco, CA, USA; 2010. pp. 1975-1981.

[37] Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Miami, FL, USA; 2009. pp. 935-942.

[38] Tharwat A. Classification assessment methods. Applied Computing and Informatics 2018.

[39] Hu Y, Zhang Y, Davis L. Unsupervised abnormal crowd activity detection using semiparametric scan statistic. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops; Portland, OR, USA; 2013. pp. 767-774.

[40] Li W, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence 2014; 36(1): 18-32.

[41] Zhu X, Liu J, Wang J, Li C, Lu H. Sparse representation for robust abnormality detection in crowded scenes. Pattern Recognition 2014; 47(5): 1791-1799.

[42] Biswas S, Babu RV. Real time anomaly detection in H. 264 compressed videos. In: Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG); Jodhpur, India; 2013. pp. 1-4.

[43] Adam A, Rivlin E, Shimshoni I, Reinitz D. Robust real-time unusual event detection using multiple xed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence 2008; 30(3): 555-560.

[44] Kim J, Grauman K. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Miami, USA; 2009. pp. 2921-2928.