IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Background subtraction in dynamic scenes using the dynamic principal component analysis

ISSN 1751-9659
Received on 26th August 2018
Revised 3rd October 2019
Accepted on 28th October 2019
E-First on 28th November 2019
doi: 10.1049/iet-ipr.2018.6095
www.ietdl.org

Achraf Djerida[1], Zhonghua Zhao[1] ✉, Jiankang Zhao[1]

[1]*Department of Instrument Science and Engineering, Shanghai Jiao Tong University, Shanghai, People's Republic of China*

✉ *E-mail: zhaozh@sjtu.edu.cn*

Abstract: This study presents a foreground detection method capable of robustly estimating the background under the presence of dynamic effects. The key contribution of this study is the use of the dynamic principal component analysis to model the serial correlation between successive frames and construct a robust pixel-based background model. The frames are normalised in hue, saturation and value colour space to reduce the effect of illumination changes. To restrict the background model, kernel density estimation is used to identify the distribution of the background time-lagged data matrix and then confidence interval limits are used to determine the corresponding detection thresholds. The foreground is detected using background subtraction. This method is tested on several common sequences such as CDnet 2014, ETSI 2014 and MULTIVISION 2013. The authors also hold comparisons based on quantitative metrics with several state-of-the-art methods. Experimental results show that their method outperforms some state-of-the-art methods and has comparable performance with some depth-based methods.

1 Introduction

Background subtraction is the process of separating a static environment (background) from a moving structure (foreground) [1]. Recently, there are numerous applications of this topic such as on surveillance and human–machine interaction [2]. A foreground detection algorithm for real-time multimedia communication systems is proposed in [3]. It avoids computationally heavy operations to make it suitable for practical applications. To boost the performance of surveillance segmentation applications, an evaluation method that takes into account detection failures and false alarms is proposed in [4]. The different proposed criteria can help the user to get the best algorithm. In [5], a machine vision algorithm, which is capable of detecting and counting fish is developed. Similar applications take place in robotics as in [6], where a visual system was developed to track moving objects.

So far, the background estimation problem has been tackled with a variety of methods [2, 7]. A simple estimation algorithm for stationary background is proposed in [8] based on running average algorithms. Although it has good computation speed, it breaks down under dynamic scenes. To detect people and interpret their behaviour, a real-time system was developed in [9]. It achieves good performance in terms of the detection accuracy; however, it is based on the assumption that the scene is less dynamic than the user. A texture-based method was used in [10] to detect moving objects. It models each pixel as a group of adaptive local histograms using a circular region. The method requires stationary camera, and it has many parameters, which can limit its performance on practical applications. To gain from the immunity of local texture features to illumination changes, photometric invariant colour measurement is proposed in [11], and it shows a robust performance on both rich texture and uniform regions. To motion and appearance. In addition to colour features, depth features have been introduced on many algorithms as in [15] provided that the background and foreground have different depths.

In this paper, we propose a novel framework for background estimation under dynamic effects using the dynamic PCA. The work is motivated by the idea of modelling a sequence of images as an output of a dynamic system. This leads to taking into account the serial correlation between the successive frames instead of assuming samples independence. PCA-based foreground detection methods show promising performance on many applications [2]; however, since PCA ignores the time correlation between samples, it initiates false estimations when the time correlation between samples cannot be ignored. In this work, we use the dynamic PCA to model the dynamic properties between successive frames to construct a robust background model. To reduce the effect of luminance variation, we normalise the images on the hue, saturation and value (HSV) colour space. Kernel density estimation is used to model the distribution of the time-lagged data matrix and foreground detection is acquired using a background subtraction scheme. To validate our method, we hold experiments on several recent sequences and with several state-of-the-art methods.

2 Related work

Background subtraction is usually done in two steps: background modelling and foreground extraction [16]. So far, this topic has been exploited widely due to its impact on different applications [17], which include advanced statistical, fuzzy, robust subspace and transform domain models [2].

Statistical models have an advantage toward illumination effects and dynamic backgrounds [2]. A real-time video

model values, and it does not need a fine-tuning algorithm. A non-parametric background model was developed in [21] to make efficient use of past values. Its detection thresholds are extended to dynamic per-pixel state variables together with the notion of dynamic controllers. A mixture of Gaussians is one of the common techniques in the field of foreground detection as in [22], where a constrained mixture of Gaussians is developed to learn the variance of the pixels as a part from the background model.

Fuzzy models have been incorporated on many background estimation methods to tackle their uncertainties [2]. A Type-2 fuzzy mixture of Gaussians model is proposed in [23] to take into account the uncertainties resulted from dynamic effects. This scheme was shown to be effective against waving trees and water rippling. To attenuate colour variations, a fuzzy colour histogram was developed in [24]. The background pixels are identified based on a similarity measure.

Similarly, a fuzzy logic model was developed in [25] based on scene parameters. It provides an automatic way for foreground detection, and it does not need user intervention. To tackle the effect of dynamic backgrounds, a fuzzy membership transformation was developed in [26] to create a rich fuzzy vector. The fuzzy statistical textural features are used as the basis for foreground detection. Experimental results show that these features can be applied to other vision-based applications.

Robust subspace methods aim at the decomposition of background and foreground via a robust subspace model [2]. To solve the background estimation problem under multiple dynamic effects, the background modelling is seen as pursuing subspaces within the video bricks as in [27], and an autoregressive (ARX) moving average model is used to characterise their appearance consistency. To make the algorithm robust to disturbances and scene changes, the subspaces are incrementally updated. Another way of factoring the background and foreground can be accomplished based on near-separable non-negative matrix factorisation as in [28], where the developed method has many advantages such as scalability and noise tolerance. Since robust PCA is the crux of many background subtraction methods, many related topics have been questioned in the literature. In [29], a study that concerns the recovery of each component when the data matrix is the superposition of a low-rank component and a sparse component. Similarly in [30] where the recovery of low-rank matrix from a high-dimensional matrix becomes a challenge when small noise and gross sparse errors exist.

Transform-based methods aim at separating the background and foreground on a different domain [2]. Representing dynamic background as a dynamic system, which has frequency responses can be a good model as shown in [31]. Under this case, the present frame is assumed to be correlated with past frames. Similarly, a frequency decomposition model that explicitly represents the scene dynamics is proposed in [32]. It can recognise some dynamics such as waves and plants motions. To identify such motion, frequency coefficients are calculated for each pixel in moving windows. To take into account both the accuracy and stability of foreground segmentation, a Gaussian mixture model is used with Walsh transform in [33] to characterise the background pixels. The Hadamard transform is used in [34] instead of the discrete cosine transform to increase the estimation speed. Experimental results record ten times faster than some of the related methods.

Recently, the work of background subtraction is directed more toward developing robust algorithms against multiple dynamic effects. In [17], an algorithm was developed to be adapted to different motion speeds using a neighbour-based intensity correction, which compares the current and previous frames to determine the adaptation decision. Foreground detection is accomplished using a threshold detection scheme based on the Otsu method. Although it shows its effectiveness under many dynamic effects, the computation burden and sensitivity toward repetitive motions decrease its performance. To tackle fast dynamic backgrounds, a region-based approach is used to generate the background model as in [35]. Experimental results show that the inclusion of neighbouring pixels improves the mixture of Gaussians model greatly. Although this scheme is suitable for dynamic scenes, many related topics still open such as the determination of the optimal region size. To achieve effective change detection, variable weights are used to scale the different samples as in [36], where it can reduce the false updating. To handle the situation where different challenges are present, a robust encoder–decoder neural network was developed in [37] to extract multi-scale features which can characterise the background robustly. A mapping is then learnt using convolutional neural networks on the decoder part. Once the model is trained on few samples, it can be used to segment the foreground on complex situations. To improve the accuracy of the segmented foreground and provide accurate ground-truth results for video surveillance tasks, a multi-resolution convolutional neural network with cascaded architecture was used in [38] to model the background. As redundant information of background and foreground is present through subsequent frames, only few samples are needed to initialise the model. To avoid parameter tuning and feature engineering, a foreground segmentation method based on convolutional neural networks is developed in [39], where the network parameters are learnt directly from the ground-truth data. To reduce the computation cost while maintaining good accuracy, a background model based on sub-superpixels is developed in [40]. As the method uses simple linear clustering techniques for model construction, it leads to more efficient performance. To boost the accuracy of both the background and foreground, a background model based on fuzzy histograms is developed in [41]. In this model, the temporal characteristics of the pixels are described using the fuzzy c-means algorithm, and an adaptive threshold is used to segment the foreground. Table 1 summarises the common background estimation categories and their basic ideas based on the work done in [2].

3 Background subtraction using dynamic PCA

3.1 Method overview

The proposed method aims at estimating the background model, which can be used to find the foreground based on a modal violation scheme. The images are modelled as an autocorrelation process to cope with the evolution of new images, which usually have a relationship with the previous samples. This formulation is motivated by the similarity to dynamic systems. The dynamic PCA is applied to the model initialisation phase to provide a mapping for each pixel. On the new feature space, statistical thresholds are developed to restrict the background. To make the proposed method robust toward illumination changes, the HSV colour space is used to normalise the images' luminance as a preprocessing phase.

Table 1 Summary of common background subtraction methods

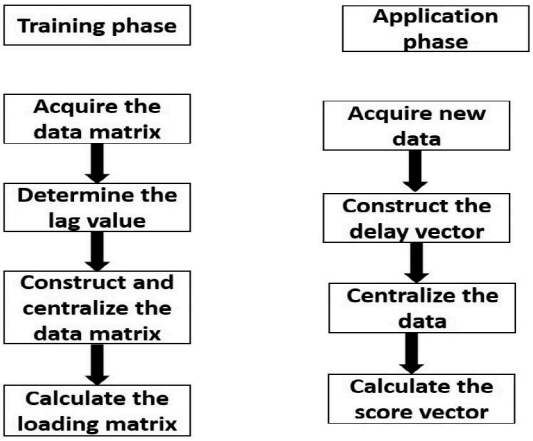

Fig. 1 *Basic steps to construct the dynamic PCA model*

3.2 Establishment of the autocorrelation model

Background estimation can be a challenging task when dynamic effects are present. The modelling of a static background is an effective method to detect any foreground provided that the estimated model is accurate [35]. In this section, we show how the dynamic PCA model can be a feasible solution to tackle such dynamic effects.

Different statistical based methods have been applied to estimate the background models such as PCA, where they show great potential in separating the background and foreground pixels [2]. Since PCA does not take into account the series correlation, the estimated model can diverge under dynamic conditions [44]. Motivated by this fact, we use dynamic PCA to estimate the background and foreground pixels. Let us assume that the data matrix $\boldsymbol{X} \in R^{n*m}$, which represents n frames with m variables, is given by

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \dots & x_{1m} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nm} \end{bmatrix} \tag{1}$$

where x_{ij} represents an intensity value.

We assume that the process is stationary [45], which means that the joint cumulative function of the following finite-order distributions are the same for every N, t_1, t_2, t_N and time shift τ:

$$\begin{aligned} &\boldsymbol{X}(t_1),\ \boldsymbol{X}(t_2),\ \dots,\ \boldsymbol{X}(t_N) \\ &\text{and} \\ &\boldsymbol{X}(t_1+\tau),\ \boldsymbol{X}(t_2+\tau),\ \dots,\ \boldsymbol{X}(t_N+\tau) \end{aligned} \tag{2}$$

To model the autocorrelation between current and previous observations, the ARX models can be used for single-input–single-output system as

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_h y_{t-h} + \beta_0 u_t + \cdots + \beta_h u_{t-h} + e_t \tag{3}$$

where y_t, u_t represent the output and input of the dynamic system at the instant t. α_i and β_i are process coefficients. e_t is a white noise process and h is a lag value.

The current output value y_t does not depend only on the current

$$\boldsymbol{\varphi} = \begin{bmatrix} y_t & u_t & \dots & y_{t-h} & u_{t-h} \\ y_{t-1} & u_{t-1} & \dots & y_{t-h-1} & u_{t-h-1} \\ \dots & \dots & \dots & \dots & \\ y_{t+h-n} & u_{t+h-n} & \dots & y_{t-n} & u_{t-n} \end{bmatrix} \tag{5}$$

The $\boldsymbol{\varphi}$ matrix reveals that the first column is linearly related to the other columns. Applying PCA to (5) is what is referred to as the dynamic PCA [46]. It involves two steps: generating the trajectory matrix and extracting the loading matrix using PCA. The trajectory matrix can be formulated based on the data matrix in (1) as [44]

$$\boldsymbol{X}'\boldsymbol{X}(h) = \begin{bmatrix} \boldsymbol{X}(1,\ \vdots\ ) & \boldsymbol{X}(2,\ \vdots\ )\cdots & \boldsymbol{X}(h,\ \vdots\ ) \\ \boldsymbol{X}(2,\ \vdots\ ) & \boldsymbol{X}(3,\ \vdots\ )\cdots & \boldsymbol{X}(h+1,\ \vdots\ ) \\ \cdot & \cdot & \cdot \\ \boldsymbol{X}(n-h+1,\ \vdots\ ) & \boldsymbol{X}(n-h+2,\ \vdots\ )\cdots & \boldsymbol{X}(n,\ \vdots\ ) \end{bmatrix} \tag{6}$$

where $\boldsymbol{X}(i,\ \vdots\ )$ means the ith row of the matrix $\boldsymbol{X}$.

To extract the loading matrix $\boldsymbol{P}$, we first centralise the trajectory matrix using its mean $\boldsymbol{\mu}$ defined as

$$\boldsymbol{\mu} = [\mu_1 \mu_2 \dots \mu_h] \tag{7}$$

Owing to the stationarity property defined by (2)

$$\mu_1 = \mu_2 = \cdots \mu_h \tag{8}$$

Since practically these conditions may not be fully satisfied, any violation of this property induces false detections. Similarly to PCA, we search for the loading matrix $\boldsymbol{P}$, which is orthogonal ($\boldsymbol{P}^T.\boldsymbol{P} = \boldsymbol{I}$) and its variances should be maximal in the directions of the new data T given by

$$T = \boldsymbol{X}'_{\mathrm{c}}.\boldsymbol{P} \tag{9}$$

where $\boldsymbol{X}'_{\mathrm{c}}$ is the centralised trajectory matrix.

To find the vectors of $\boldsymbol{P}$, an optimisation problem has to be solved. To maximise the variance of the scores, the following expression is maximised:

$$\underset{j}{\operatorname{argmax}}\, t_j^T t_j = \underset{j}{\operatorname{argmax}} \boldsymbol{p}_j^T x^T x \boldsymbol{p}_j \tag{10}$$

where t_j corresponds to the jth component of the score matrix $\boldsymbol{T}$, $\boldsymbol{p}_j$ corresponds to the jth component of the loading matrix $\boldsymbol{P}$ and $\boldsymbol{x}$ is a vector from the trajectory matrix $\boldsymbol{X}'_{\mathrm{c}}$.

Solving this problem using Lagrange multipliers leads to [47]

$$[A - \lambda \boldsymbol{x}] \times \boldsymbol{p}_j = 0 \tag{11}$$

This is a classical eigenvalue problem, where $A = \boldsymbol{X}'^{T}_{\mathrm{c}}\boldsymbol{X}'_{\mathrm{c}}$, λ represents its eigenvalues and $\boldsymbol{p}_j$ represents its eigenvectors. Fig. 1 shows the basic steps for the construction of the dynamic PCA

the stationarity property defined by (2). To solve this problem, there are mainly two methods:

- *Threshold adaptation:* In this method, the thresholds are updated by comparing the current and previous images and based on a change detector the background is updated. This method can be a good solution for slow variations; however, it can get into trouble under great changes.
- *Model normalisation:* This method looks for an illumination invariant space where the background model is independent of these changes. In this work, we adopt this scheme since it can be more reliable under high illumination effects.

To reduce the illumination effects, we transform the red, green and blue (RGB) image to the HSV colour space, where the normalisation takes place. The reason for selecting this colour space for normalisation is owed to the characteristics of its channels, which can lead to the separation of the luminance from the chrominance (colour) [48]. While the hue of a colour determines which pure colour it resembles, its saturation determines the amount of grey, which is contained in colour. The value channel measures the intensity of the colour.

Consider a test image I in the RGB space, and we first convert it to the HSV space HSV (I) to calculate the mean of the value channel as

$$a_{\mathrm{v}} = \frac{1}{m_1 \,.\, m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} V \tag{12}$$

where $\boldsymbol{V}$ represents the value channel and m_1 and m_2 are the number of rows and columns. To normalise images, we centralise the value channel using the following equation:

$$\boldsymbol{V}' = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} a_{\mathrm{d}} + \alpha(\boldsymbol{V}_{ij} - a_{\mathrm{v}}) \tag{13}$$

where a_{d} represents the desired mean value and α is a constant used to modify the contrast of the image.

The value of a_{d} is determined during the background initialisation and all the images are normalised using (13). Fig. 2 shows the effects of a_{d} and α on a reference image after converting it back to the RGB colour space.

Let the background initialisation stage runs over $i = 1 \cdots n$ frames. The value of a_{d} is determined as the average value of the calculated mean using (12). To make the estimation process more robust, we assign for each pixel with image coordinates I_{ij} a dynamic PCA model M_{ij} generated by (6) and (9). The $\boldsymbol{X}$ matrix defined by (1) is used to represent the values of the RGB channels as

$$\boldsymbol{X}_{ij} = \begin{bmatrix} I_{ij}(1,\,1) & I_{ij}(1,\,2) & I_{ij}(1,\,3) \\ \cdot & \cdot & \cdot \\ I_{ij}(n,\,1 & I_{ij}(n,\,2) & I_{ij}(n,\,3) \end{bmatrix} \tag{14}$$

where $I_{ij}(p,\,q)$ represents the intensity of the channel q at the frame p.

Although PCA is used generally for dimension reduction, it can be used to decorrelate the data and provide a monitoring statistic to characterise the background [46]. This is in coincidence with many

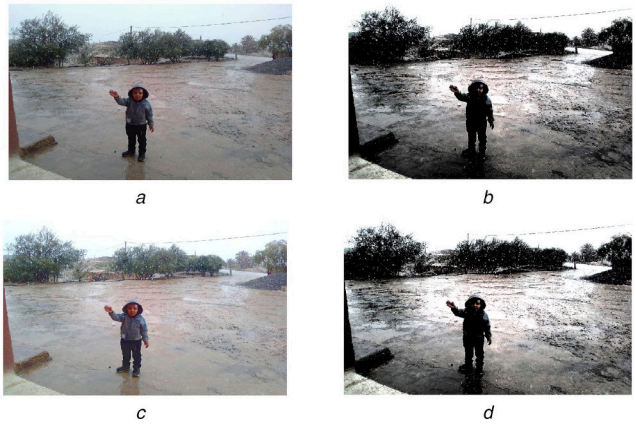

Fig. 2 *Effect of the HSV colour space coefficients*
(a) a_d = reference, $\alpha = 1$, ***(b)*** a_d = reference, $\alpha = 4$, ***(c)*** $a_{\mathrm{d}} = 0.70$, $\alpha = 1$, ***(d)*** $a_{\mathrm{d}} = 0.70$, $\alpha = 4$

Clearly, for the processing of a new frame, we need to save the last h samples in the delay vector as

$$\boldsymbol{V}_d(n) = [\boldsymbol{X}_{ij}(n-h+1,\,\vdots\,)\boldsymbol{X}_{ij}(n-h+2,\,\vdots\,)\cdots\boldsymbol{X}_{ij}(n,\,\vdots\,)] \tag{16}$$

After the centralisation of the trajectory matrix, the score space is generated by (9) as

$$\boldsymbol{T}_{ij} = \boldsymbol{X}'_{ij} \,.\, \boldsymbol{P}_{ij} \tag{17}$$

where $\boldsymbol{P}_{ij}$ is the loading matrix for pixel $(i,\,j)$.

To determine the background model on the score space, we use the norm of the score vectors as a statistical measure as

$$\mathbf{Norm}_{ij}(l) = \sqrt{\sum_{k=1}^{3.h} . \, T_{ij}(k)^2} \tag{18}$$

where l denotes the frame number.

To restrict the background region in the score space, upper and lower thresholds are determined. Since assuming that the norm statistic follows Gaussian distribution can be violated practically, we use kernel density estimation to find its distribution as a sum of kernels [51]:

$$D(s) = \frac{1}{n \,.\, \sigma} \sum_{l=1}^{n} K\left(\frac{s - \mathbf{Norm}_{ij}(l)}{\sigma}\right) \tag{19}$$

where s is a data point, which we want to find its density value $D(s)$, $\mathbf{Norm}_{ij}(l)$ is the norm of the pixel $(i,\,j)$ at the frame l, K is a kernel function (normal density function), σ is a smoothing parameter and n represents the total number of initialisation frames.

To find the confidence interval thresholds, we calculate the cumulative distribution function $C(s)$ as

$$C(s) = \int_{-\infty}^{s} D(t) \,.\, \mathrm{d}t \tag{20}$$

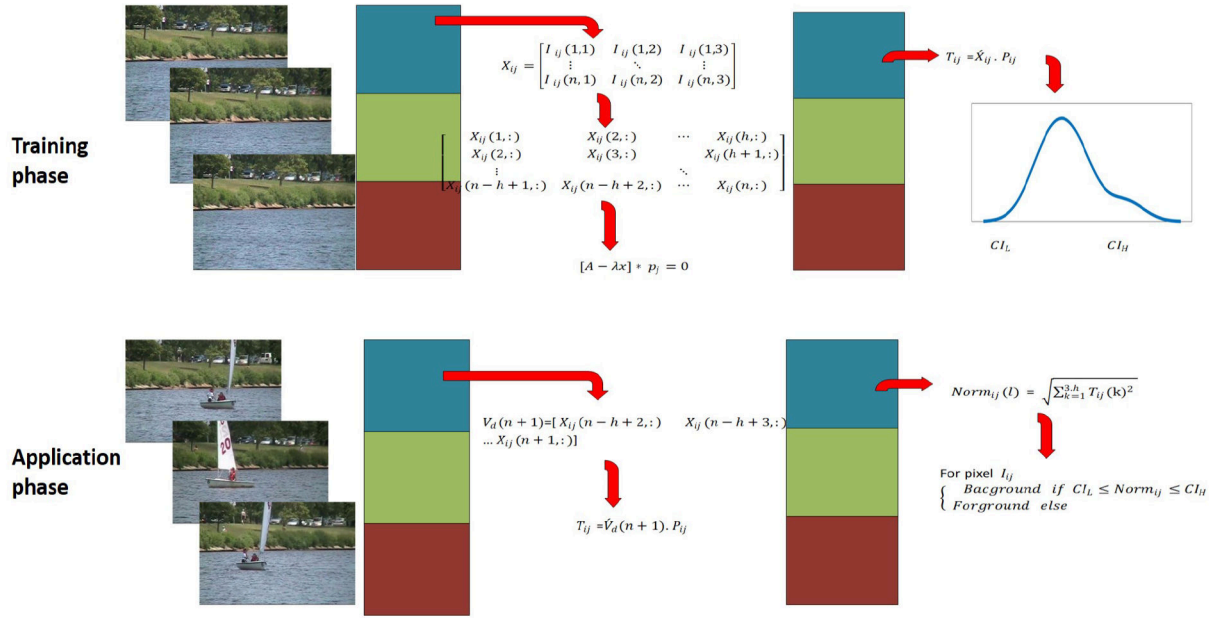

Fig. 3 *Basic steps for the construction of the background model*

Sequence name	Sample frame	Image size	Number of frames
Boats		320*240	7999
Overpass		320*240	3000
Sofa		320*240	2750
Boulevard		352*240	2500
Office		360*240	2050
Pedestrians		360*240	1099
CopyMachine		720*480	3400
ChairBox		640*480	529
GenSeq1		640*480	410
DCamSeq2		640*480	670
Fall01cam1		640*480	160
ColorCam2		640*480	360
Topviewlab2		640*480	650

is defined by (18). Each pixel is classified based on the following constraints.

For pixel I_{ij}

$$\begin{matrix} \text{background} & \text{if} & \text{CI}_\text{L} \leq \mathbf{Norm}_{ij} \leq \text{CI}_\text{H} \\ \text{foreground} & & \text{else} \end{matrix} \tag{22}$$

4 Experimental results and discussion

4.1 Datasets and evaluation criteria

In this section, we use 13 videos to evaluate the proposed method with comparisons with the state-of-the-art methods. We select videos from four recent datasets: the CDNet2014 [52], ETSI [53], MULTIVISION [54] and UR fall detection [55]. Fig. 4 reports some characteristics of these videos:

(i) *ETSI:* It is one of the first published RGB depth (RGBD) benchmarks for evaluating background estimation methods under a variety of effects [53]. From this dataset, we select four videos from four categories: GenSeq1 from the illumination changes, DCamSeq2 from depth camouflage, TopViewLab2 from the out of range sensor and ColorCam2 from colour Camouflage.
(ii) *MULTIVISION:* It is characterised by the inclusion of the depth information in addition to the RGB images [54]. From this dataset, we select the ChairBox sequence, which includes illumination effects.
(iii) *UR fall detection:* It contains different fall scenarios with both RGB and depth images. We select Fall01cam1 sequence, which contains shadow effects.
(iv) *CDnet 2014:* It is an expansion to CDnet 2012, and it is created to evaluate change and motion detection approaches [52]. We select seven videos from five categories: boats and overpass from dynamic background, copy machine from shadows, boulevard from camera jitter, sofa from intermittent motion and office and pedestrians from baseline.

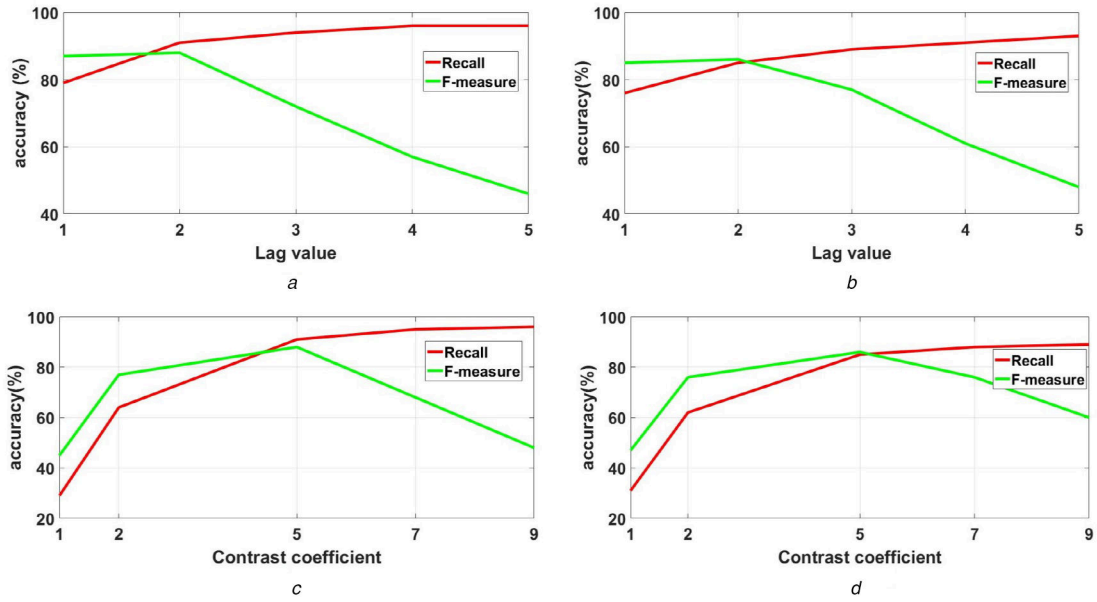

Fig. 5 *Effect of the lag value and contrast coefficient on some sequences*
(a) Effect of the lag value as applied to pedestrians, ***(b)*** Effect of the lag value as applied to Overpass, ***(c)*** Effect of the contrast coefficient as applied to pedestrians, ***(d)*** Effect of the contrast coefficient as applied to overpass

Table 2 Recall comparisons of the CDnet2014

Methods	AAPSA	CP3	BMOG	WeSam	Proposed
boat	0.63	0.84	0.74	0.48	0.79
overpass	0.70	0.67	0.99	0.59	0.84
pedestrian	0.96	0.90	0.98	0.95	0.89
sofa	0.54	0.82	0.55	0.64	0.72
boulevard	0.54	0.77	0.93	0.58	0.81
office	0.93	0.94	0.55	0.89	0.90
copy machine	0.91	0.88	0.55	0.88	0.90
average	0.74	0.83	0.76	0.72	0.84
standard deviation	0.19	0.09	0.21	0.19	0.07

4.2 Effect of the proposed algorithm parameters

In this section, we determine the values of the lag value h and the contrast coefficient α based on their effect on the recall and F-measure. Since the recall measures the detection rate and the F-measure takes into account both the recall and precision measures, both of them can give a good evaluation of the quality of estimation. In determining the lag value, the contrast coefficient is held constant, and then the reverse is true for the contrast coefficient.

In the literature, there is no optimal way to determine the lag value h [46]; however, usually some heuristic methods are used. Common methods include the use of 1 or 2 lagging samples as in fault detection systems [46] or making it a fraction (25%) from the total number of observation [44]. In this work, we determine its value as the one, which results in the best performance based on the recall and F-measure. For this reason, we choose the pedestrians and overpass as test sequences as shown in Fig. 5. We can see that both the recall and F-measure increase as the lag value increases until value 2, where the recall continues increasing while the F-measure starts dropping. On both sequences, we see that the until five, where it starts to decrease rapidly. The increase of the contrast coefficient can improve the algorithm performance provided that it is less than its breaking point which is in this case 5. Therefore, the contrast coefficient is fixed to 4.5 just before the breaking point to avoid possible degradations.

4.3 Qualitative and quantitative results

In this section, we hold experiments on different datasets with comparisons with some recent state-of-the-art methods. The parameters of the algorithm are fixed for all the datasets as follows: the number of frames used for background initialisation $N = 30$, the lag value $h = 2$ and the contrast coefficient $\alpha = 4.5$. To make fair comparisons, we compare our algorithm with some methods, which have been applied to these datasets and reported by their authors. The resulted images of the state-of-the-art methods for the CDnet 2014 are reported in the dataset website [52] and for the rest of datasets are provided on the scene background modeling (SBM)-RGBD website [57]. To evaluate the three criteria for all the methods, we use the provided scripts in [52].

Table 3 F-measure comparisons of the CDnet2014

Methods	AAPSA	CP3	BMOG	WeSam	Proposed
boat	0.76	0.54	0.84	0.64	0.85
overpass	0.82	0.77	0.96	0.72	0.87
pedestrian	0.96	0.94	0.92	0.96	0.88
sofa	0.69	0.83	0.63	0.76	0.80
boulevard	0.63	0.47	0.58	0.72	0.78
office	0.95	0.96	0.63	0.93	0.93
copy machine	0.76	0.86	0.64	0.92	0.86
average	0.80	0.77	0.74	0.81	0.85
standard deviation	0.12	0.19	0.16	0.13	0.05

Table 4 Precision comparisons of the CDnet2014

Methods	AAPSA	CP3	BMOG	WeSam	Proposed
boat	0.98	0.40	0.96	0.96	0.94
overpass	0.98	0.91	0.94	0.93	0.90
pedestrian	0.97	0.97	0.87	0.96	0.88
sofa	0.94	0.85	0.75	0.94	0.89
boulevard	0.75	0.34	0.43	0.93	0.76
office	0.98	0.99	0.74	0.98	0.97
copy machine	0.66	0.83	0.77	0.97	0.82
average	0.89	0.76	0.78	0.95	0.88
standard deviation	0.13	0.27	0.18	0.02	0.07

The boat video contains dynamic background and moving foreground (boats), which can make high accuracy challenging. The proposed algorithm achieves the second-best score (87%) after BMOG (96%) on the overpass sequence, whereas WeSam gets the least score (72%). Similar to the boat, overpass contains dynamic background represented by tree waving. Pedestrians and office contain mild changes of shadow, camera jitter, dynamic background and intermittent motion. Under such circumstances, the proposed algorithm achieves the least score (88%) on the pedestrian sequence, which is not too far from the best score (96%) obtained by AAPSA and WeSam. On the Office sequence, our detector achieves the third-best score (93%). Under intermittent effects on the sofa sequence, our algorithm gets the second-best score (80%) after CP3 (83%). Considering the camera jitter effects on the Boulevard video, the proposed method produces the best score (78%) followed by WeSam (72%), whereas CP3 gets the least score (50%). This good performance continues under shadow effects as shown in the copy machine, where our algorithm gets the second-best score (86%) after WeSam (92%).

If we rank the methods in terms of their F-measure, our method produces the best average (85%) followed by WeSam (81%) and AAPSA (80%). BMOG shows the least score (74%). Our method and BMOG show the best robustness toward dynamic background, whereas WeSam and CP3 methods have the best performance when different effects are applied simultaneously. This can be deduced from the scatter measure, where our algorithm has the minimum standard deviation (0.05) compared with the maximum (0.19) obtained by CP3.

Similar to the F-measure, the precision and recall criteria show that our method advances all the methods in terms of the average recall and produces the third-best score in terms of precision. The proposed algorithm shows good balance between the detection and precision through almost all the videos. This effect can be foreground. For boat sequence, we can see the effect of dynamic background on the results of BMOG and CP3, where some scattered pixels have been detected falsely as foreground. For WeSam and AAPSA, only portion of the boat has been detected. Our method shows good robustness to the dynamic background even though a small part from the boat has not been detected. This performance has been confirmed on overpass, where the person has been almost missed by WeSam, and some pixels have been falsely declared as foreground by AAPSA and BMOG. In boulevard, one of the cars is missed by AAPSA and WeSam. The proposed method detects both cars with almost a complete shape. For copy machine, one of the two persons has been missed by BMOG while some pixels have been detected falsely as foreground by AAPSA and CP3. The shadow effects have been successfully ignored by most of the methods. For office, all the methods show good detection and precision, except for a small part from the book which has been missed by CP3 and a small part from the person which has been missed by BMOG. Similarly, for pedestrians where some pixels have been missed by WeSam; however, the rest of the methods show complete detection. For sofa, two objects are completely missed by BMOG, whereas one object is completely missed by AAPSA; however, the proposed method detects all the three objects.

4.3.2 Results on ETSI, MULTIVISION and UR fall detection sequences: In this section, we compare the proposed method with three state-of-the-art methods: RPCA [61], CwisardH + [62] and RGB-self-organized background subtraction (SOBS) [63]. The RPCA and CwisardH + use depth information, which can allow us to evaluate the proposed algorithm better, which relies only on colour information.

Tables 5–7 show comparisons between these methods based on the recall, precision and F-measure. In terms of the average F-

Fig. 6 *Qualitative results for some methods on CDnet2014*

Table 5 Recall comparisons of the ETSI, MULTIVISION and UR datasets

Methods	SRPCA	RGB-SOBS	CwisardH+	Proposed
ChairBox	0.92	0.77	0.88	0.73
GenSeq1	1	0.98	1	0.94
DCamSeq2	0.79	0.99	0.43	0.93
Fall01cam1	0.84	0.98	0.79	0.86
ColorCam2	1	0.23	1	0.72
Topviewlab2	0.98	0.91	0.95	0.84
standard deviation	0.09	0.30	0.22	0.09
average	0.92	0.81	0.84	0.84

Table 6 F-measure comparisons of the ETSI, MULTIVISION and UR datasets

Methods	SRPCA	RGB-SOBS	CwisardH+	Proposed
ChairBox	0.85	0.86	0.92	0.83
GenSeq1	0.93	0.95	0.91	0.89
DCamSeq2	0.81	0.91	0.57	0.74
Fall01cam1	0.75	0.96	0.84	0.88
ColorCam2	0.88	0.36	0.97	0.80
Topviewlab2	0.80	0.91	0.92	0.82
standard deviation	0.06	0.23	0.15	0.06
average	0.84	0.83	0.86	0.83

Table 7 Precision comparisons of the ETSI, MULTIVISION and UR datasets

Methods	SRPCA	RGB-SOBS	CwisardH +	Proposed
ChairBox	0.80	0.98	0.97	0.95
GenSeq1	0.87	0.92	0.83	0.84
DCamSeq2	0.82	0.84	0.84	0.61
Fall01cam1	0.67	0.95	0.91	0.89
ColorCam2	0.79	0.77	0.95	0.91
Topviewlab2	0.68	0.91	0.90	0.81
standard deviation	0.08	0.08	0.06	0.12

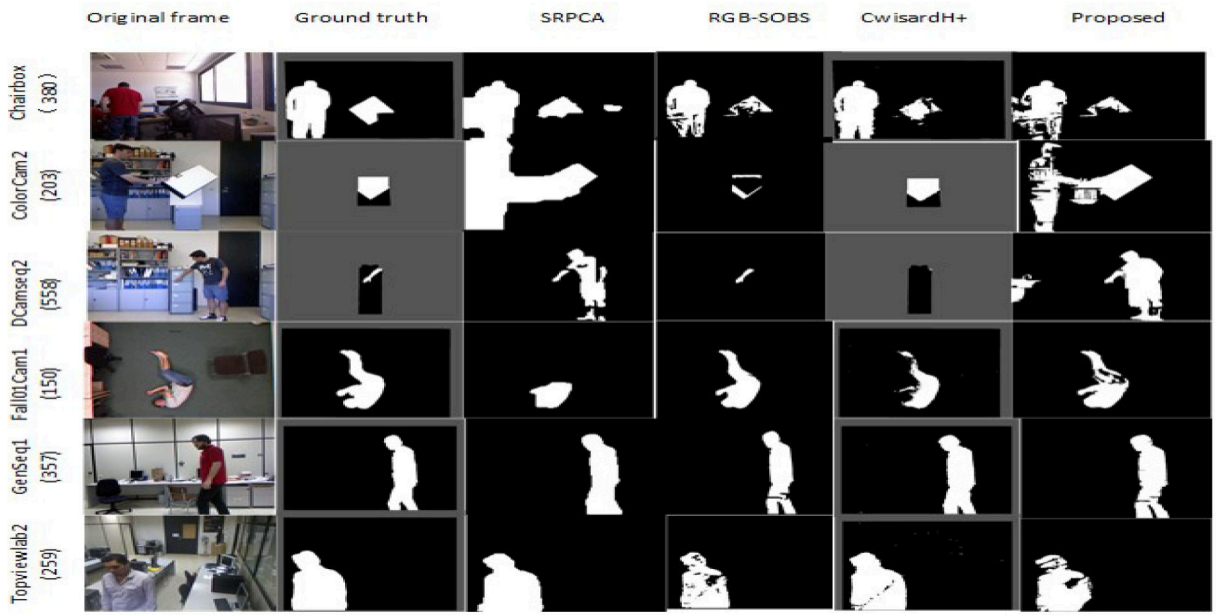

Fig. 7 *Qualitative results for some different methods on ETSI, MULTIVISION and UR fall detection datasets*

Table 8 Approximate cost of some methods

Methods	Resolution	Speed, FPS
proposed	320×240	2
WeSam	320×240	2
CwisardH +	640×480	2
AAPSA	240×360	20
CP3	320×240	20
BMOG	320×240	102
SRPCA	320×240	4
RGB-SOBS	640×480	4

RGB-SOBS, where it fails to detect a large part from this paper. Similarly on DCamseq2, where the foreground and background have the same depth, the proposed method detects almost the person's hand in contrast to CwisardH + which fails to detect it completely. For Fall01Cam1 sequence, our method successfully separates the foreground from the background, and it misses only a small part from the person in contrast to SRPCA which misses almost half the foreground size. On GenSeq1 and Topviewlab2, all the methods show good overall segmentation of the foreground. This section highlights the importance of depth features at situations, where the background and foreground do not have the same depth.

4.4 Complexity analysis

To analyse the complexity of the proposed algorithm, we examine the basic blocks of modelling, detection and state update steps based on the worst-case analysis and big-O notation. As our method is pixel based, during the model initialisation (training) three nested loops are executed, where two loops concern the frame dimensions and one loop for the number of training frames. Therefore, the computation complexity during the training stage follows $O(n^3)$, where n denotes the size of each loop. Once the shows comparisons based on approximate frames per second (FPS). The results of the state-of-the-art methods are taken from [52, 57]. We can see that our method has comparable speed with WeSam and SRPCA, which have a computation cost around 2 and 4 FPSs, respectively. This value is the result of a pixel-based approach which is used by our method to model the background. Although BMOG has a high speed (102 FPS) its performance compared with our method, as seen from previous tables, is lower with around 10% in terms of the average F-measure.

4.6 Discussion

By examining the results on the previous tables and figures, we can draw several important remarks. Although the performance of the proposed method is not much better than the state-of-the-art methods, it has good characteristics as it maintains reasonable scores under severe conditions in contrast to some methods, where their scores reduce to <50%. This advantage can be confirmed by the reported standard deviations on each dataset. The second advantage can be extracted by examining the qualitative results, where we remark that the proposed method maintains reasonable segmentations under severe cases in contrast to some methods, where they fail under some conditions to preserve the basic shape of the foreground. Although the proposed method suffers from its high computation cost due to the use of a pixel-based approach, it is a challenge as shown in Table 8 for many background subtraction methods to maintain robust performance with fast computation time. These remarks highlight the potential of the proposed method to have better performance provided that both the computation complexity and segmentation accuracy are enhanced in a way that improves the model robustness while the algorithm complexity reduces to a linear function.

5 Future work

As we pointed in the discussion, the proposed method has the potential to be further enhanced, so that both computation time and

initialisation method, the robustness of the model is boosted knowing that background initialisation can be performed at situations, where all the frames have foreground objects.

The computation complexity in our work can be enhanced by using the fact that each pixel has an independent dynamic PCA model from the neighbouring pixels. Hence, instead of processing the pixels serially as we did in our implementation, parallel processing can yield faster computation time. If high computation time is more important for an application, the proposed method can be adjusted to have just one global dynamic PCA model with local thresholds for each pixel to determine its properties. By using just one dynamic PCA model, the whole algorithm can be implemented as few matrix summations and multiplications, which can boost greatly the computation time.

6 Conclusion

In this paper, we propose a robust background subtraction method for dynamic scenes. The use of dynamic PCA permits to model the serial correlation between successive frames. Our method shows good performance compared with state-of-the-art methods in terms of the three common criteria: recall, precision and F-measure. It results in a robust model toward dynamic backgrounds, camera jitter, intermittent motion and other dynamic effects. Furthermore, comparisons with depth-based methods show superior performance in terms of the standard deviation across different sequences and show comparable performance in terms of the average accuracy. Future work will consider the improvement of computation speed and the use of other features such as depth information.

7 References

[1] Zeng, D., Zhu, M.: 'Background subtraction using multiscale fully convolutional network'. *IEEE Access*, 2018, **2018**, pp. 16010–16021

[2] Bouwmans, T.: 'Traditional and recent approaches in background modeling for foreground detection: an overview', *Comput. Sci. Rev.*, 2014, **11**, pp. 31–66

[3] Chien, S.-Y., Ma, S.-Y., Chen, L.-G.: 'Efficient moving object segmentation algorithm using background registration technique', *IEEE Trans. Circuits Syst. Video Technol.*, 2002, **12**, (7), pp. 577–586

[4] Nascimento, J.C., Marques, J.S.: 'Performance evaluation of object detection algorithms for video surveillance', *IEEE Trans. Multimed.*, 2006, **8**, (4), pp. 761–774

[5] Spampinato, C., Chen-Burger, Y.-H., Nadarajan, G., *et al.*: 'Detecting, tracking and counting fish in low quality unconstrained underwater videos'. *VISAPP*, 2008, **1**, (2), pp. 514–519

[6] Watanabe, Y., Fabiani, P., Le Besnerais, G.: 'Simultaneous visual target tracking and navigation in a GPS-denied environment'. Int. Conf. Advanced Robotics 2009 ICAR 2009, Munich, Germany, 2009, pp. 1–6

[7] Choudhury, S.K., Sa, P.K., Bakshi, S., *et al.*: 'An evaluation of background subtraction for object detection vis-a-vis mitigating challenging scenarios', *IEEE Access*, 2016, **4**, pp. 6133–6150

[8] Lai, A.H., Yung, N.H.: 'A fast and accurate scoreboard algorithm for estimating stationary backgrounds in an image sequence'. Int. Symp. IEEE 1998 Proc. Circuits and Systems 1998 ISCAS'98, Monterey, USA, vol. **4**, 1998, pp. 241–244

[9] Wren, C., Azarbayejani, A., Darrell, T., *et al.*: 'P nder: real-time tracking of the human body. media lab 353'. 1995

[10] Heikkila, M., Pietikainen, M.: 'A texture-based method for modeling the background and detecting moving objects', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, (4), pp. 657–662

[11] Zeng, D., Zhu, M., Xu, F., *et al.*: 'Extended scale invariant local binary pattern for background subtraction', *IET Image Process.*, 2018, **12**, (8), pp. 1292–1302

[12] Marghes, C., Bouwmans, T., Vasiu, R.: 'Background modeling and foreground detection via a reconstructive and discriminative subspace learning approach'. Proc. Int. Conf. Image Processing, Computer Vision, and Pattern Recognition (IPCV) The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), Las Vegas, USA, 2012, p. 1

[18] Mukherjee, D., Jonathan Wu, Q.: 'Real-time video segmentation using Student's *t* mixture model', *Procedia Comput. Sci.*, 2012, **10**, pp. 153–160

[19] Ding, J., Li, M., Huang, K., *et al.*: 'Modeling complex scenes for accurate moving objects segmentation'. Asian Conf. Computer Vision, Springer, 2010, pp. 82–94

[20] Barnich, O., Van Droogenbroeck, M.: 'Vibe: a powerful random technique to estimate the background in video sequences'. IEEE Int. Conf. Acoustics, Speech and Signal Processing 2009 ICASSP 2009, Taipai, Taiwan, 2009, pp. 945–948

[21] Hofmann, M., Tiefenbacher, P., Rigoll, G.: 'Background segmentation with feedback: the pixel-based adaptive segmenter'. 2012 IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, USA, 2012, pp. 38–43

[22] Sheri, A.M., Rafique, M.A., Jeon, M., *et al.*: 'Background subtraction using Gaussian–Bernoulli restricted Boltzmann machine', *IET Image Process.*, 2018, **12**, (9), pp. 1646–1654

[23] El Baf, F., Bouwmans, T., Vachon, B.: 'Type-2 fuzzy mixture of Gaussians model: application to background modeling'. Int. Symp. Visual Computing, Springer, 2008, pp. 772–781

[24] Kim, W., Kim, C.: 'Background subtraction for dynamic texture scenes using fuzzy color histograms', *IEEE Signal Process. Lett.*, 2012, **19**, (3), pp. 127–130

[25] Rosell-Ortega, J., Garcia-Andreu, G., Rodas-Jorda, A., *et al.*: 'A combined self-configuring method for object tracking in colour video'. 2010 20th Int. Conf. Pattern Recognition (ICPR), Istanbul, Turkey, 2010, pp. 2081–2084

[26] Chiranjeevi, P., Sengupta, S.: 'New fuzzy texture features for robust detection of moving objects', *IEEE Signal Process. Lett.*, 2012, **19**, (10), pp. 603–606

[27] Lin, L., Xu, Y., Liang, X., *et al.*: 'Complex background subtraction by pursuing dynamic spatio-temporal models', *IEEE Trans. Image Process.*, 2014, **23**, (7), pp. 3191–3202

[28] Kumar, A., Sindhwani, V.: 'Near-separable non-negative matrix factorization with *l*1 and Bregman loss functions'. Proc. 2015 SIAM Int. Conf. Data Mining SIAM, Vancouver,Canada, 2015, pp. 343–351

[29] Candès, E.J., Li, X., Ma, Y., *et al.*: 'Robust principal component analysis?', *J. ACM*, 2011, **58**, (3), p. 1

[30] Zhou, Z., Li, X., Wright, J., *et al.*: 'Stable principal component pursuit'. 2010 IEEE Int. Symp. Information Theory Proc. (ISIT), Austin, USA, 2010, pp. 1518–1522

[31] Wren, C.R., Porikli, F.: 'Waviz: spectral similarity for object detection'. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, Cambridge, Massachusetts, USA, 2005, pp. 55–61

[32] Porikli, F., Wren, C.: 'Change detection by frequency decomposition: wave-back'. Proc. Workshop on Image Analysis for Multimedia Interactive Services, Cambridge, Massachusetts, USA, 2005

[33] Tezuka, H., Nishitani, T.: 'A precise and stable foreground segmentation using fine-to-coarse approach in transform domain'. 15th IEEE Int. Conf. Image Processing 2008 ICIP 2008, San Diego, USA, 2008, pp. 2732–2735

[34] Baltieri, D., Vezzani, R., Cucchiara, R.: 'Fast background initialization with recursive Hadamard transform'. 2010 Seventh IEEE Int. Conf. Advanced Video and Signal based Surveillance (AVSS), Boston,USA, 2010, pp. 165–171

[35] Varadarajan, S., Miller, P., Zhou, H.: 'Region-based mixture of Gaussians modelling for foreground detection in dynamic scenes', *Pattern Recognit.*, 2015, **48**, (11), pp. 3488–3503

[36] Jiang, S., Lu, X.: 'WeSamBe: a weight-sample-based method for background subtraction', *IEEE Trans. Circuits Syst. Video Technol.*, 2018, **28**, (9), pp. 2105–2115

[37] Lim, L.A., Keles, H.Y.: 'Foreground segmentation using convolutional neural networks for multiscale feature encoding', *Pattern Recognit. Lett.*, 2108, **112**, pp. 256–262

[38] Wang, Y., Luo, Z., Jodoin, P.-M.: 'Interactive deep learning method for segmenting moving objects', *Pattern Recognit. Lett.*, 2017, **96**, pp. 66–75

[39] Babaee, M., Dinh, D.T., Rigoll, G.: 'A deep convolutional neural network for video sequence background subtraction', *Pattern Recognit.*, 2018, **76**, pp. 635–649

[40] Chen, Y.-Q., Sun, Z.-L., Lam, K.: 'An effective sub-superpixel-based approach for background subtraction', *IEEE Trans. Ind. Electron.*, 2020, **67**, (1), pp. 601–609

[41] Yu, T., Yang, J., Lu, W.: 'Dynamic background subtraction using histograms based on fuzzy c-means clustering and fuzzy nearness degree', *IEEE Access*, 2019, **7**, pp. 14671–14679

[42] Oliver, N.M., Rosario, B., Pentland, A.P.: 'A Bayesian computer vision system for modeling human interactions', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (8), pp. 831–843

[43] Yamazaki, M., Xu, G., Chen, Y.-W.: 'Detection of moving objects by independent component analysis'. Asian Conf. Computer Vision, Springer, 2006, pp. 467–478

[44] Mina, J., Verde, C.: 'Fault detection using dynamic principal component

[49] Howley, T., Madden, M.G., O'Connell, M.-L., *et al.*: 'The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data', *Knowl.-Based Syst.*, 2006, **19**, (5), pp. 363–370
[50] Popelínskỳ, L.: 'Combining the principal components method with different learning algorithms'. Proc. ECML/PKDD IDDM Workshop (Integrating Aspects of Data Mining, Decision Support and Meta-Learning 2001, Pennsylvania, USA, 2000
[51] Silverman, B.W.: '*Density estimation for statistics and data analysis*' (Routledge, USA, 2018)
[52] Goyette, N., Jodoin, P.-M., Porikli, F., *et al.*: 'changedetection.net: A new change detection benchmark dataset'. 2012 IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, USA, 2012, pp. 1–8
[53] Camplani, M., Salgado, L.: 'Background foreground segmentation with RGB-D kinect data: an efficient combination of classifiers', *J. Vis. Commun. Image Represent.*, 2014, **25**, (1), pp. 122–136
[54] Fernandez-Sanchez, E.J., Diaz, J., Ros, E.: 'Background subtraction based on color and depth using active sensors', *Sensors*, 2013, **13**, (7), pp. 8895–8915
[55] Kwolek, B., Kepski, M.: 'Human fall detection on embedded platform using depth maps and wireless accelerometer', *Comput. Methods Programs Biomed.*, 2014, **117**, (3), pp. 489–501
[56] Hassan, M.A., Malik, A.S., Saad, N., *et al.*: 'Evaluation metric for rate of background detection'. Instrumentation and Measurement Technology Conf. Proc. (I2MTC) 2016 IEEE Int., Taipei, Taiwan, 2016, pp. 1–5
[57] Camplani, M., Maddalena, L., Alcover, G.M., *et al.*: 'A benchmarking framework for background subtraction in RGBD videos'. Int. Conf. Image Analysis and Processing, Springer, Catania, Italy, 2017, pp. 219–229
[58] Martins, I., Carvalho, P., Corte-Real, L., *et al.*: 'BMOG: boosted Gaussian mixture model with controlled complexity'. Iberian Conf. Pattern Recognition and Image Analysis, Springer, Faro, Portugal, 2017, pp. 50–57
[59] Ramírez-Alonso, G., Chacón-Murguía, M.I.: 'Auto-adaptive parallel SOM architecture with a modular analysis for dynamic object segmentation in videos', *Neurocomputing*, 2016, **175**, pp. 990–1000
[60] Liang, D., Hashimoto, M., Iwata, K., *et al.*: 'Co-occurrence probability-based pixel pairs background model for robust object detection in dynamic scenes', *Pattern Recognit.*, 2015, **48**, (4), pp. 1374–1390
[61] Javed, S., Bouwmans, T., Sultana, M., *et al.*: 'Moving object detection on RGB-D videos using graph regularized spatiotemporal RPCA'. Int. Conf. Image Analysis and Processing, Springer, Catania, Italy, 2017, pp. 230–241
[62] De Gregorio, M., Giordano, M.: 'CwisardH: background detection in RGBD videos by learning of weightless neural networks'. Int. Conf. Image Analysis and Processing, Springer, Catania, Italy, 2017, pp. 242–253
[63] Maddalena, L., Petrosino, A.: 'The SOBS algorithm: what are the limits?'. 2012 IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, USA, 2012, pp. 21–26
[64] Djerida, A., Zhao, Z., Zhao, J.: 'Robust background generation based on an effective frames selection method and an efficient background estimation procedure (FSBE)', *Signal Process., Image Commun.*, 2019, **78**, pp. 21–31