Dynamic Codebook for Foreground Segmentation in a Video

Worapan Kusakunniran [1], Non-member and **Rawitas Krungkaew** [2], Member

ABSTRACT

The foreground segmentation in a video is a way to extract changes in image sequences. It is a key task in an early stage of many applications in the computer vision area. The information of changes in the scene must be segmented before any further analysis could be taken place. However, it remains with difficulties caused by several real-world challenges such as cluttered backgrounds, changes of the illumination, shadows, and long-term scene changes. This paper proposes a novel method, namely a dynamic codebook (DCB), to address such challenges of variations in the background scene. It relies on a dynamic modeling of the background scene. Initially, a codebook is constructed to represent the background information of each pixel over a period of time. Then, a dynamic boundary of the codebook will be made to support variations of the background. The revised codebook will always be adaptive to the new background's environments. This makes the foreground segmentation more robust to the changes of background scene. The proposed method has been evaluated by using the changedetection.net (CDnet) benchmark which is a well-known video dataset for testing change-detection algorithms. The experimental results and comprehensive comparisons have shown a very promising performance of the proposed method.

Keywords: Foreground, Segmentation, Video, Dynamic, Codebook

1. INTRODUCTION

The foreground segmentation aims to detect moving objects in a video sequence. It is a very important step in many computer vision based applications [1][2][3][4]. For example, in [1], to analyze the performance of a tennis match, key events such as the tennis ball being hit and the tennis ball bouncing on the ground must be detected beforehand. In order to detect such events, the foreground of the tennis ball must be extracted. This is where the foreground segmentation will be employed. It is similar to the work in [2] related to the application domain of the traffic surveillance, the foreground of the vehicle must be segmented before any further analysis such as the count of vehicles, the calculation of average speed of individual vehicle will be taken place. In [3][4], the foreground or the movement of human must be extracted before being used in recognizing human gaits and human actions respectively.

In addition, there are many techniques [5][6][7][8][9][10][11][12][13][14][15][16][17][18][19][20] proposed to address the problem of the foreground segmentation in a video. For example, the method [5], namely frame differencing, is one of the most conventional techniques for the foreground segmentation. It works well under the static background scene especially in the indoor environment. The core concept is to detect any movement in a scene by subtracting the current frame from the previous frame. However, this technique cannot extract a full silhouette of a moving object. It can detect only a moving part of an object. Also, it has a difficulty in detecting an object that moves with a very slow speed such as a movement of a few pixels per dozen frames [6].

In [16], the performance of the foreground segmentation is boosted by using the artificial neural network to learn variations in the background scene. The learning rate must be calibrated to match the nature of each video scene. It is automatically chosen by using the fuzzy technique and the coherence-based self-organizing background subtraction algorithm. In [17], the genetic programming is used to combine the outputs of multiple foreground-detection algorithms. The proposed technique will match the right algorithm for the right video. As a result, it can achieve the high performance based on the CDnet benchmark. However, several challenges of the foreground segmentation still remain unsolved.

In [11], a Gaussian mixture model (GMM) is used to model the background information in each pixel of the video frame. This is performed under the assumption that the GMM can model a variation of background's intensity values in each particular pixel by using a multimodal background model. Then, the likelihood of a pixel's value being a part of backgrounds or foregrounds will be determined based on a simple heuristic of the constructed GMM of that particular pixel. The one that does not fit to the constructed GMM is identified as a foreground pixel.

Manuscript received on July 31, 2016 ; revised on December 13, 2016.

Manuscript received on January 15, 2017.

[1] The author is with Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom, Thailand 73170, Phone(+662)4410909, E-mail: worapan.kunmahidol.ac.th

[2] The author is with Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom, Thailand 73170, Phone(+662)4410909, E-mail: rawitagmail.com

Otherwise, it will be identified as a background pixel.

The key advantage of this technique is that it can deal with repetitive motions in background scenes. The multimodal background model of the GMM can be used to represent multiple surfaces appeared in each particular pixel such as water surfaces. However, it cannot easily and efficiently model fast variations in background scenes with just a few Gaussians. Moreover, this technique also has to deal with the trade-off issue on adjusting the learning rate. For the low learning rate, the background model will be gradually adjusted to the recent background scene. In this case, it will have a difficulty in adapting the background model to the sudden change in the background scene. In contrast, for the high learning rate, the background model will be rapidly adjusted to the recent background scene. It will make the slow-moving object to be eventually absorbed into the background model and faultily classified as a part of the background [12].

To overcome such limitation of the GMM-based method, the codebook-based technique was adopted for the foreground segmentation in a video by not making use of any parametric parameters [12]. One codebook will be constructed to model the background information of each pixel. It will consist of one or more codewords. One codeword will represent one variation of the background model in the scene. For the background scene with the high variations, the large number of codewords will be created for each codebook during the training phase.

Then, in the testing phase, if a pixel's value does not belong to any codewords of the corresponding codebook, then that pixel will be identified as the foreground. The advantage of this technique is that it can efficiently deal with a dynamic background scene. However, it is not flexible enough to deal with a very high dynamic background environment. Also, it can suffer from the false negative foreground detection due to its quantization criterion [13]. In [13][14][15], the codebook-based method was improved by combining with other techniques.

Based on our literature review, the challenge of variations in the background scene is remaining unsolved. It can be caused by the moving of the background scene itself such as waving trees and water surfaces, or caused by the unstable camera. In this paper, the DCB-based method is proposed to address such challenge. The boundary of the codeword will be adjusted and adapted to the content of the recent background scene. In addition, the new codeword can be created to represent a new background information. Also, the existing codeword can be deleted if it represents the out-of-date background information

It has been known that the evaluation is the crucial process to identify the strength and weakness of each method [19]. In this paper, the proposed method has been evaluated based on the datasets and evaluation approaches as described in the CDnet [18]. The experimental results and comprehensive comparisons are carried out.

It can be concluded that the proposed method outperforms the existing techniques in the literature review, for the case of the baseline category containing the combination of mild challenges on the foreground segmentation in a video, which are a subtle background motion, isolated shadows, an abandoned object, and stopping pedestrians for a short period. Particularly, the proposed method is also shown to be very promising on solving the challenges of the unstable camera in the camera jitter category and the strong background motion in the dynamic background category.

The rest of this paper is organized as follows. Section 2 describes the details of the proposed method. Section 3 shows the experimental results and the relevant discussions. Then, the conclusion is drawn in section 4.

2. THE PROPOSED DYNAMIC CODEBOOK (DCB)

The proposed DCB is explained in this section. The DCB itself is for the background modeling. Later, its models can be used for the foreground segmentation in a video, which will be also explained in the section 2.2 It is developed to overcome the limitation of the conventional codebook-based method [21] in order to address the challenge of the dynamic background scene. The conventional codebook-based method performs under the assumption that the cylindrical codebook-model in RGB color space can sufficiently model the pixel intensity distortions. However, for the high-dynamic background scene, such variations of the background information cannot be coped by using the cylindrical shape. To address this problem, the dynamic adjustment of the codeword's range/shape is embedded in the development process of DCB.

2.1 Codebook

The framework of the codebook-based method is shown in the algorithm 1. One codebook is constructed as the background model of each pixel. It contains at least one codeword. Each codeword is represented by a range of pixel's intensity values of the background of that pixel. The number of codewords for each codebook depends on a degree of variations in the background scene of the training phase. The high variation in the background pixel will result in the high number of codewords for the corresponding codebook. Thus, the codebook is created by observing changes of pixel's intensity in video frames.

In this paper, to simplify the explanation, it is assumed that the codebook is constructed for a particular pixel i (CB_i). Each codeword j ($CW_{i,j}$) of CB_i is modeled by using the rgb vector $v_{i,j} = (r_{i,j}, g_{i,j}, b_{i,j})$

Algorithm 1 Construction of the codebook (CB_i) in the training phase

Input: The given video sequence of the background scene (V)
Output: The codebook (CB_i)

```
Initialize CB_i to be empty (i.e. contains no codeword)
for (each frame t in V )
{
    Initialize flag = codeword not found
    for (each codeword CW_{i,j})
    {
        if(δ(I_{i,t}, v_{i,j}) < ε and I_min < I_{i,t} < I_max)
        {
            flag = codeword found
        {
    }
    if (flag = codeword found)
    {
        v_{i,j} = ((f×r_{i,j}+r)/(f+1), (f×g_{i,j}+g)/(f+1), (f×b_{i,j}+b)/(f+1))
        h_{i,j} = (min(I_{i,t}, I_min), max(I_{i,t}, I_max), f + 1, neg, p, t)
    {
    else
    {
        The new codeword for CB_i is constructed
    {
}
return CB_i
```

and the hex-tuple $h_{i,j} = (I_{min}, I_{max}, f, neg, p, q)$, where I_{min} is the minimum brightness value assigned to $CW_{i,j}$, I_{max} is the maximum brightness value assigned to $CW_{i,j}$, f is the frequency or the number of times that $CW_{i,j}$ occurs, neg is the maximum negative run length (MNRL) which is the longest period that $CW_{i,j}$ is not accessed, and p and q are the first and the last access times of $CW_{i,j}$ respectively. In addition, the color space does not have to always be rgb. It can be other color spaces such as $l\alpha\beta$ as used in this paper.

In the training phase, a video sequence of the background scene is used to construct codebooks as the background models. If the training data sufficiently covers most of possible background variations, then the constructed codebooks will be robust to such variations during the process of the foreground segmentation.

The training process of constructing the codebook (CB_i) begins with checking the pixel's intensity value $I_{i,t} = (r, g, b)$ from each background frame at time t against each existing $CW_{i,j}$ based on the following two criteria. First, as shown in the equation (1), the color distortion (δ) between $I_{i,t}$ and $CW_{i,j}$ is less than the detection threshold (ϵ).

$$\delta(I_{i,t}, v_{i,j}) < \epsilon \tag{1}$$

The color distortion is calculated as shown in the equation (2).

$$\delta(I_{i,t}, v_{i,j}) = \sqrt{||I_{i,t}||^2(1 - cos^2\theta)} \tag{2}$$

where $||I_{i,t}||$ is the $l2$-norm of $I_{i,t}$ as shown in the equation (3) and $cos\theta$ is the cosine similarity between $I_{i,t}$ and $v_{i,j}$ as shown in the equation (4). As in the calculation, $cos\theta$ is used to measure the similarity between $I_{i,t}$ and $v_{i,j}$. Thus, 1 - $cos^2\theta$ is used to measure the difference or the distortion between $I_{i,t}$ and $v_{i,j}$. The $||I_{i,t}||^2$ is used as the multiplication factor in the equation (2), in order to geometrically normalize the codeword's vector ($v_{i,j}$) to the brightness of the input pixel ($I_{i,t}$) [21].

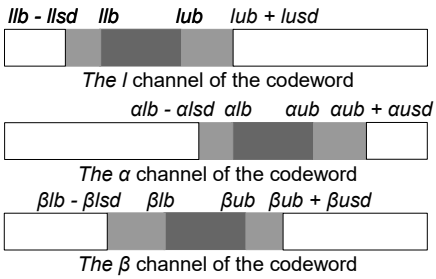

Fig.1: *The structure of each codeword in DCB.*

$$||I_{i,t}|| = \sqrt{r^2 + g^2 + b^2} \tag{3}$$

$$cos\theta = \frac{r \times r_{i,j} + g \times g_{i,j} + b \times b_{i,j}}{\sqrt{r^2 + g^2 + b^2}\sqrt{r_{i,j}^2 + g_{i,j}^2 + b_{i,j}^2}} \tag{4}$$

Second, the brightness of the input pixel is between the minimum and maximum brightness values. If there is at least one $CW_{i,j}$ that satisfies these two criteria, then the corresponding $v_{i,j}$ and $h_{i,j}$ will be updated as shown in the equations (5) and (6). Otherwise, the new codeword will be created for this CB_i by using the information of the input pixel.

In the testing phase, these two criteria are used for segmenting the foreground from the background. That is, if there is at least one $CW_{i,j}$ that satisfies the two criteria, the pixel will be classified as the background. Also, the corresponding $v_{i,j}$ and $h_{i,j}$ will be updated as shown in the equations (5) and (6). Otherwise, the pixel will be classified as the foreground.

$$v_{i,j} = (\frac{f \times r_{i,j} + r}{f + 1}, \frac{f \times g_{i,j} + g}{f + 1}, \frac{f \times b_{i,j} + b}{f + 1}) \tag{5}$$

$$h_{i,j} = (min(I_{i,t}, I_{min}), max(I_{i,t}, I_{max}), f+1, neg, p, t) \tag{6}$$

2.2 Dynamic codebook (DCB)

In this paper, the DCB is developed under the $l\alpha\beta$ color space. The l channel represents the illumination information, while α and β channels represent the color information. This color space was shown to be an effective choice for the process of foreground segmentation in a video [22]. This is because, in the $l\alpha\beta$ color space, the chromatic component (i.e. α and β) is almost completely separated from the achromatic component (i.e. l) [23].

Table 1: *The symbols used in the structure of each codeword in DCB.*

Symbols	Descriptions
l	the pixel's intensity value in the l channel
α	the pixel's intensity value in the α channel
β	the pixel's intensity value in the β channel
llb	the lower bound of a codeword in the l channel
αlb	the lower bound of a codeword in the α channel
βlb	the lower bound of a codeword in the β channel
lub	the upper bound of a codeword in the l channel
αub	the upper bound of a codeword in the α channel
βub	the upper bound of a codeword in the β channel
$llsd$	the standard deviation of $\|llb - l\|$ over a training period
αlsd	the standard deviation of $\|\alpha lb - \alpha\|$ over a training period
βlsd	the standard deviation of $\|\beta lb - \beta\|$ over a training period
$lusd$	the standard deviation of $\|lub - l\|$ over a training period
αusd	the standard deviation of $\|\alpha ub - \alpha\|$ over a training period
βusd	the standard deviation of $\|\beta ub - \beta\|$ over a training period

Moreover, in the proposed DCB, the l channel of the codeword can be modeled to be more dynamic than the α and β channels of the codeword. This will make the DCB to be more robust to changes of illumination in the background scene, when compared with the rgb color space.

As mentioned, DCB has the key feature of the dynamic decision boundary of each codeword. Thus, the range and/or shape of the codeword in the $l\alpha\beta$ color space will be dynamically adjusted to the recent context of the background environment. To simplify the explanation, Fig. 1 illustrates the structure of each codeword and Table 1 describes the meaning of symbols used in DCB as shown in Fig. 1.

The structure of each codeword consists of three channels according to the used color space i.e. $l\alpha\beta$. In each channel of the codeword, there are two layers of the boundary in order to cope with variations of the background in each pixel. The first layer is ranged by using the lower bound and upper bound. The second layer is extended from the first layer by using the standard deviation of the variations of the background.

Table 2: *The construction of a new codeword in DCB. Note: δ is a constant value for a new codeword construction.*

Components	Values
llb	$l - \delta$
lub	$l + \delta$
$llsd$	δ
$lusd$	δ
αlb	$\alpha - \delta$
αub	$\alpha + \delta$
αlsd	δ
αusd	δ
βlb	$\beta - \delta$
βub	$\beta + \delta$
βlsd	δ
βusd	δ

Table 3: *The seven measurements for the performance evaluation.*

Measurements	Calculations
Re	$\frac{TP}{TP+FN}$
Sp	$\frac{TN}{TN+FP}$
FPR	$\frac{FP}{FP+TN}$
FNR	$\frac{FN}{TP+FN}$
PWC	$\frac{100(FN+FP)}{TP+FN+FP+TN}$
Pr	$\frac{TP}{TP+FP}$
F1	$\frac{2(Pr)(Re)}{Pr+Re}$

This proposed DCB-based method has two phases which are training and testing phases. They are described in the following paragraphs.

2.2.1 The training phase

It needs a video sequence of the background scene for constructing the codebooks CB_i. For DCB, the pixel' values $I_{i,t} = (l, \alpha, \beta)$ from each background frame at time t will be checked against each existing $CW_{i,j}$ based on the following three conditions.

$$(llb - llsd) \leq l \leq (lub + lusd) \tag{7}$$

Algorithm 2 The updates of the matched codeword $CW_{i,j}$ in DCB

Input: The pixel' values $I_{i,t} = (l, \alpha, \beta)$ from each background frame at time t
Output: The updated codeword $CW_{i,j}$

```
if (llb − llsd) ≤ l ≤ (lub + lusd)
{
    if (l < llb)
    {
        llb = llb − llsd
        llsd = ((llsd × f) + k(llb − l)) / (f + 1)
    }
    if (l > lub)
    {
        lub = lub + lusd
        lusd = ((lusd × f) + k(l − lub)) / (f + 1)
    }
}
if (αlb − αlsd) ≤ α ≤ (αub + αusd)
{
    if (α < αlb)
    {
        αlb = αlb − αlsd
        αlsd = ((αlsd × f) + k(αlb − α)) / (f + 1)
    }
    if (α > αub)
    {
        αub = αub + αusd
        αusd = ((αusd × f) + k(α − αub)) / (f + 1)
    }
}
if (βlb − βlsd) ≤ β ≤ (βub + βusd)
{
    if (β < βlb)
    {
        βlb = βlb − βlsd
        βlsd = ((βlsd × f) + k(βlb − β)) / (f + 1)
    }
    if (β > βub)
    {
        βub = βub + βusd
        βusd = ((βusd × f) + k(β − βub)) / (f + 1)
    }
}
return CW_{i,j}
```

$$(\alpha lb - \alpha lsd) \leq \alpha \leq (\alpha ub + \alpha usd) \tag{8}$$

$$(\beta lb - \beta lsd) \leq \beta \leq (\beta ub + \beta usd) \tag{9}$$

If there is no codeword that satisfies with these three conditions, a new codeword of CB_i will be created as shown in Table 2. Otherwise, the matched codeword will be updated as shown in the algorithm 2. In the Table 2, the constant value (δ) for the construction of a new codeword is empirically adjusted. In this paper, it is set to be 5.

2.2.2 The testing phase

In the testing phase, for each pixel i in a current frame, if it belongs to any codeword of CB_i by satisfying the three conditions in the equations (7), (8) and (9), then it will be classified as the background pixel. Otherwise, it will be classified as the foreground pixel. If the pixel is classified as the background, it will be used to update the matched codeword as shown in the algorithm 2.

In the algorithm 2, k is the learning factor for updating the value of the standard deviation of the pixel's intensity value away from the corresponding codeword. In this paper, k is empirically set to be 1. Therefore, the codewords and codebooks modeling the background will be dynamically updated based on the changes of the scene. That means, the DCB has a property to deal with variations of the background in a cluttered environment.

In addition, the out-of-date codeword can be deleted from the corresponding codebook. In this case, the out-of-date codeword means the codeword in which there is no any pixels belong to it for a period of time which can be empirically adjusted. In this paper, this period of time is set to be 50 frames. However, the out-of-date codeword will not deleted if it is the only one codeword left in the corresponding codebook.

3. EXPERIMENTAL RESULTS

In this paper, the CDnet benchmark [18] is used to evaluate the performance of the proposed method. The results of the foreground segmentation by using the proposed method are compared with the ground truth provided by the CDnet. The performance of the proposed method is also compared to the performances of the other existing methods in the literature review, using the same CDnet dataset. The CDnet dataset provides realistic, camera capture videos (with no CGI) from both indoor and outdoor environments with various camera ranging. The dataset of the six categories provides about 70,000 frames in 31 videos. The six categories are baseline, dynamic background, camera jitter, shadows, intermittent object motion and thermal.

The baseline category contains a combination of mild challenges on the foreground segmentation in a video, as can be seen in Fig. 2. The challenges include a subtle background motion, isolated shadows, an abandoned object, and stopping pedestrians for a short period. The dynamic background category contains a strong background motion such as a shimmering water and a shaking tree. The camera jitter category contains videos recorded by unstable cameras. The intermittent object motion category contains scenes of topping pedestrians for a short period or suddenly start moving pedestrians. The shadows category contains strong shadows and the thermal category contains videos recorded by far-infrared cameras.

This paper focuses on the baseline dataset. This is because the proposed method has been developed for the normal scenes and the scenes with mild to medium variations caused by several challenges as mentioned above for the baseline category. While the other five categories focus on the strong varia-

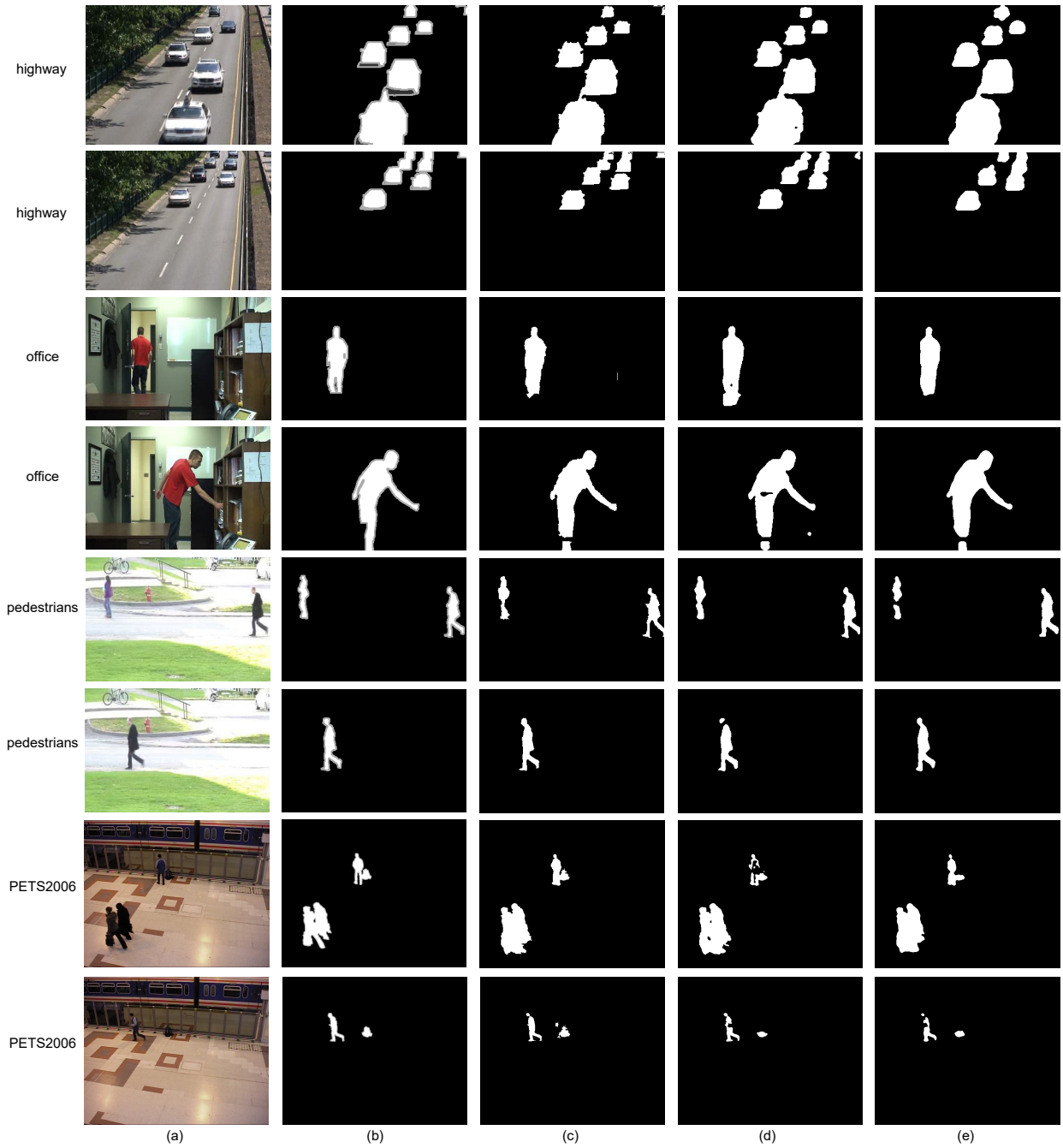

Fig.2: *The sample segmented foreground images by using the proposed method, the IUTIS-1 [17] and the multimode background subtraction [24], compared with the ground truth images. (a) Original images. (b) Ground truth images. (c) Segmented images by using the proposed method. (d) Segmented images by using the IUTIS-1 [17]. (e) Segmented images by using the multimode background subtraction [24].*

tion of its only type. However, the additional experiments and comparisons are also carried out under these five categories of the dataset, in order to preliminarily analyze the results of applying the proposed method on the strong variations. Moreover, the proposed method will be further developed in the future work to cope with such strong variations.

The seven measurements are used to calculate the average ranking and performance across all methods in the CDnet. As described in Table 3, they are Recall (Re), Specificity (Sp), False Positive Rate (FPR), False Negative Rate (FNR), Percentage of Wrong Classifications (PWC), Precision (Pr), and F-Measure (F1). These seven measurements are calculated based on the four measurements of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

The TP, TN, FP and FN are calculated by comparing the segmentation result based on the proposed method with the ground truth. TP is calculated based on the foreground pixels that are correctly identified as the foreground. TN is calculated based on

Table 4: *The performance comparisons between the proposed method and the other existing methods in the literature review, based on the baseline video category from the CDnet benchmark.*

Method	**Average Re**	**Average Sp**	**Average FPR**	**Average FNR**	**Average PWC**	**Average F1**	**Average Pr**
Multimode Background Subtraction [24]	0.9158	0.9979	0.0021	0.0842	0.4361	0.9287	0.9431
IUTIS-1 [17]	0.9214	0.9979	0.0021	0.0786	0.4538	0.9298	0.9391
SOBS-CF [25]	0.9347	0.9978	0.0022	0.0653	0.3912	0.9299	0.9254
FTSG [26]	0.9513	0.9975	0.0025	0.0487	0.4766	0.9330	0.9170
Spectral-360 [27]	0.9616	0.9968	0.0032	0.0384	0.4265	0.9330	0.9065
M4CD Version 1.5 [28]	0.9521	0.9975	0.0025	0.0479	0.4402	0.9272	0.9057
CwisarDRP [29]	0.8580	0.9981	0.0019	0.1420	0.8778	0.8880	0.9347
CwisarDH [30]	0.8972	0.9980	0.0020	0.1028	0.5679	0.9145	0.9337
AAPSA [31]	0.9092	0.9979	0.0021	0.0908	0.5826	0.9183	0.9286
M4CD Version 1.0 [28]	0.9382	0.9976	0.0024	0.0618	0.4934	0.9204	0.9063
C-EFIC [32]	0.9455	0.9970	0.0030	0.0545	0.5201	0.9309	0.9170
KDE-ElGammal [11]	0.8969	0.9977	0.0023	0.1031	0.5499	0.9092	0.9223
EFIC [33]	0.9349	0.9971	0.0029	0.0651	0.5223	0.9172	0.9023
Mahalanobis distance [34]	0.3154	0.9991	0.0009	0.6846	2.8698	0.4642	0.9270
KNN [35]	0.7934	0.9979	0.0021	0.2066	1.2840	0.8411	0.9245
CP3-online [36]	0.8501	0.9972	0.0028	0.1499	0.7725	0.8856	0.9252
RMoG [37]	0.7082	0.9981	0.0019	0.2918	1.5935	0.7848	0.9125
AMBER [38]	0.8784	0.9973	0.0027	0.1216	0.9233	0.8813	0.8980
IUTIS-2 [17]	0.7452	0.9978	0.0022	0.2548	1.5115	0.7913	0.9100
Euclidean distance [34]	0.8385	0.9955	0.0045	0.1615	1.0260	0.8720	0.9114
GMM (Zivkovic) [39]	0.8085	0.9972	0.0028	0.1915	1.3298	0.8382	0.8993
Multiscale Spatio-Temporal BG Mode [40]	0.8137	0.9970	0.0030	0.1863	1.1478	0.8450	0.8870
GMM (Stauffer and Grimson) [41]	0.8180	0.9948	0.0052	0.1820	1.5325	0.8245	0.8461
GraphCutDiff [42]	0.7028	0.9960	0.0040	0.2972	1.9757	0.7147	0.8093
The proposed method (DCB)	0.9456	0.9975	0.0025	0.0544	0.4366	0.9359	0.9272

the background pixels that are correctly identified as the background. FP is calculated based on the foreground pixels that are incorrectly identified as the background. FN is calculated based on the background pixels that are incorrectly identified as the foreground.

The sample segmented foreground images by using the proposed method are shown in Fig. 2. They are from four different videos including highway, office, pedestrians, and PETS2006 in the CDnet benchmark. The results shown in Fig. 2 demonstrate the efficiency and robustness of the proposed method under various movements and conditions of the background scenes.

Table 5: *The precision comparisons between the proposed method and the other existing methods in the literature review, based on the other five categories from the CDnet benchmark including 1) dynamic background (DB), 2) camera jitter (CJ), 3) shadows (SD), 4) intermittent object motion (IO) and 5) thermal (TM). The average precision from the five categories (Avg1) and the average precision from the two categories of dynamic background and camera jitter (Avg2) are shown in the last two columns respectively.*

Method	**DB**	**CJ**	**IO**	**SD**	**TM**	**Avg1**	**Avg2**
Multimode Background Subtraction [24]	0.8651	0.8443	0.7827	0.3481	0.8268	0.7334	0.8547
IUTIS-1 [17]	0.3305	0.5299	0.5485	0.6032	0.9245	0.5873	0.4302
SOBS-CF [25]	0.5953	0.6405	0.5464	0.5899	0.8715	0.6487	0.6179
FTSG [26]	0.9129	0.7645	0.8512	0.5005	0.9088	0.7875	0.8387
Spectral-360 [27]	0.8456	0.8387	0.7374	0.5815	0.9114	0.7829	0.8422
CwisarDRP [29]	0.8723	0.8713	0.8543	0.5773	0.9116	0.8174	0.8718
CwisarDH [30]	0.8499	0.8516	0.7417	0.5547	0.8786	0.7753	0.8508
AAPSA [31]	0.7336	0.8021	0.7139	0.5877	0.8795	0.7434	0.7679
M4CD Version 1.0 [28]	0.6806	0.7901	0.8000	0.5590	0.9452	0.7550	0.7354
C-EFIC [32]	0.6993	0.8157	0.5823	0.4791	0.8690	0.6891	0.7575
KDE-ElGammal [11]	0.5732	0.4862	0.4609	0.6217	0.8974	0.6079	0.5297
EFIC [33]	0.6849	0.6389	0.5634	0.4846	0.849	0.6442	0.6619
Mahalanobis distance [34]	0.7451	0.8564	0.5098	0.8726	0.9932	0.7954	0.8008
KNN [35]	0.6931	0.7018	0.7121	0.3979	0.9186	0.6847	0.6975
CP3-online [36]	0.6122	0.4562	0.5631	0.5914	0.7663	0.5978	0.5342
RMoG [37]	0.7288	0.7605	0.8026	0.3097	0.9365	0.7076	0.7447
AMBER [38]	0.7990	0.8493	0.7530	0.4658	0.8514	0.7437	0.8242
IUTIS-2 [17]	0.5564	0.7184	0.8374	0.4480	0.9395	0.6999	0.6374
Euclidean distance [34]	0.4487	0.3753	0.4995	0.5763	0.8877	0.5575	0.4120
GMM (Zivkovic) [39]	0.6213	0.4872	0.6458	0.5428	0.8706	0.6335	0.5543
Multiscale Spatio-Temporal BG Mode [40]	0.5515	0.3979	0.6016	0.5282	0.8403	0.5839	0.4747
GMM (Stauffer and Grimson) [41]	0.5989	0.5126	0.6688	0.5352	0.8652	0.6361	0.5558
GraphCutDiff [42]	0.5357	0.5918	0.8315	0.4260	0.9111	0.6592	0.5638
The proposed method (DCB)	0.7632	0.9107	0.5291	0.3706	0.8502	0.6848	0.8370

As shown in Fig. 2, in the highway video, the segmentation performances of the three methods are equally good. However, the proposed method is shown to be better than the IUTIS-1 [17] and the multimode background subtraction [24] methods, in the office, pedestrians and PETS2006 videos. The proposed method can provide good quality segmented silhouettes, when compared with the results of the two methods [17][24].

In the row 3 of Fig. 2 (d), for the office video, the IUTIS-1 [17] incorrectly segments the shadow as the foreground. In the row 4 of Fig. 2 (d), for the office video, the IUTIS-1 [17] generates the incomplete silhouette.

In the row 5 of Fig. 2 (e), for the pedestrians video, the multimode background subtraction [24] generates

the incomplete silhouette. The human silhouette is cut into two parts. Similarly, in the row 6 of Fig. 2 (d), for the pedestrians video, the IUTIS-1 [17] also cuts the human silhouette into two parts.

In the rows 7 and 8 of Fig. 2 (d) and (e), it can be seen that the IUTIS-1 [17] and the multimode background subtraction [24] cannot provide the complete silhouettes.

Based on the 'highway' video, the proposed method is shown to be robust to the shadow of the trees, the small movement of the background (i.e. the moving trees), the fast moving objects (i.e. the moving cars), the small and far moving objects, and the outdoor environment. Based on the 'office' video, the proposed method is shown to be robust to the slow moving objects, and the indoor environment. Based on the 'pedestrians' video, the proposed method is shown to be robust to the variation of the daylight. It is also shown to be able to provide the clean silhouettes of the human body. Based on the 'PETS2006' video, the proposed method is shown to be able to work well under the conditions of the top camera-view. Also, it can segment the group of people and the unattended object.

In addition, the comprehensive comparisons between the performance of proposed method and the performance of the existing methods in the literature review are shown in Table 4. This is done based on the baseline video category provided by the CDnet. In Table 4, it can be seen that the proposed DCB-based method outperforms the other existing methods, in average (i.e. reported by the CDnet). This average ranking is calculated by averaging the ranks of the method in Re, Sp, FPR, FNR, PWC, Pr and F1. The proposed DCB achieves very high recall, specificity and precision of 94.56%, 99.75% and 92.72% respectively.

Table 5 shows the precision comparisons between the proposed method and the other existing methods in the literature, based on the other five categories from the CDnet benchmark including 1) dynamic background (DB), 2) camera jitter (CJ), 3) shadows (SD), 4) intermittent object motion (IO), and 5) thermal (TM). For the camera jitter category, the proposed method achieves the highest precision, when compared with the other existing methods in the literature. Its performance is also promising for the categories of dynamic background and thermal. This is because the proposed DCB can adapt quickly to the changes of the background. Multiple codewords are created to model variations of the background in each pixel of the scene, which can be caused by the camera jitter and/or the dynamic background itself. The dynamic of the codewords can make it quickly handle the variations of the background.

However, the proposed method cannot achieve the good performance for the shadow category. This is because the proposed DCB is explicitly developed to address the challenge of the variations in the background scene, not particularly for the shadow. In the future work, to handle the case of the shadow, the DCB can be modified especially in the luminance channel of the codewords, or the additional technique of the shadow removal can be applied to the segmentation results.

In Table 5, the average precision from the five categories (Avg1) and the average precision from the two categories of dynamic background and camera jitter (Avg2) are displayed in the last two columns respectively. The Avg2 is used to validate the performance of the techniques for the challenge of the variations in the background scene, which is the focus of this paper.

As shown in Table 5, based on the Avg1 and Avg2, the proposed method is shown to outperform many methods in the literature. In addition, when compared with [17][28][31][32][35], the proposed method achieves the lower precision in average from the five categories, but it can achieve the significantly higher precision in average from the two categories. When compared with [34][37][38], the proposed method can achieve the slightly higher precision in average from the two categories, however, it can perform significantly better in the case of the baseline category (as shown in Table 4). When compared with [24][26][27][29][30], the proposed method achieves the lower precisions in both cases of Avg1 and Avg2, however, it can perform better in the case of the baseline category (as shown in Table 4).

4. CONCLUSION

In this paper, the DCB-based foreground segmentation method in a video is proposed. It is developed to overcome the limitation of the conventional codebook-based method. The variations in background scenes are coped by using the dynamic boundary of each codeword in the DCB. Thus, the proposed method can be more robust under the cluttered environments in the real scenes. It can cope with the variations in the background scene. The CDnet benchmark is used to evaluated the performance of the proposed method. It has been shown that the proposed method achieves a very promising performance by outperforming other existing methods in the literature, for the case of the baseline category which contains the mild challenges of a subtle background motion, isolated shadows, an abandoned object, and stopping pedestrians for a short period. The proposed method is also shown to be very robust for the case of strong variations in the background scene caused by the unstable camera and the dynamic background in the camera jitter and dynamic background categories respectively.

References

[1] F. Yan, W. Christmas, and J. Kittler, "A tennis ball tracking algorithm for automatic annotation of tennis match," in *British machine vision conference*, vol. 2, 2005, pp. 619–628.

[2] R. A. Hadi, G. Sulong, and L. E. George, "Vehicle detection and tracking techniques: a concise review," *arXiv preprint arXiv:1410.5894*, 2014.

[3] W. Kusakunniran, "Recognizing gaits on spatio-temporal feature domain," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 9, pp. 1416–1423, 2014.

[4] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[5] N. Prabhakar, V. Vaithiyanathan, A. P. Sharma, A. Singh, and P. Singhal, "Object tracking using frame differencing and template matching," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 4, no. 24, pp. 5497–5501, 2012.

[6] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell, "Towards robust automatic traffic scene analysis in real-time," in *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, vol. 1. IEEE, 1994, pp. 126–131.

[7] J. N. Kapur, P. K. Sahoo, and A. K. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer vision, graphics, and image processing*, vol. 29, no. 3, pp. 273–285, 1985.

[8] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[9] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[10] T. Horprasert, D. Harwood, and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *Ieee iccv*, vol. 99, 1999, pp. 1–19.

[11] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *European conference on computer vision*. Springer, 2000, pp. 751–767.

[12] A. Ilyas, M. Scuturici, and S. Miguet, "Real time foreground-background segmentation using a modified codebook model," in *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*. IEEE, 2009, pp. 454–459.

[13] Y. Li, F. Chen, W. Xu, and Y. Du, "Gaussian-based codebook model for video background subtraction," in *International Conference on Natural Computation*. Springer, 2006, pp. 762–765.

[14] S. ITing, S.-C. Hsu, and C.-L. Huang, "Hybrid codebook model for foreground object segmentation and shadow/highlight removal," *Journal of Information Science and Engineering*, vol. 30, pp. 1965–1984, 2014.

[15] M. A. Mousse, E. C. Ezin, and C. Motamed, "Foreground-background segmentation based on codebook and edge detector," in *Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on*. IEEE, 2014, pp. 119–124.

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

[17] S. Bianco, G. Ciocca, and R. Schettini, "How far can you get by combining change detection algorithms?" *arXiv preprint arXiv:1505.02921*, 2015.

[18] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "Cdnet 2014: an expanded change detection benchmark dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 387–394.

[19] S. Brutzer, B. Höferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1937–1944.

[20] H. Sajid and S.-C. S. Cheung, "Background subtraction for static & moving camera," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4530–4534.

[21] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground–background segmentation using codebook model," *Real-time imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[22] P. Noriega, B. Bascle, and O. Bernier, "Local kernel color histograms for background subtraction." in *VISAPP (1)*, 2006, pp. 213–219.

[23] F. J. López-Rubio, E. López-Rubio, R. M. Luque-Baena, E. Dominguez, and E. J. Palomo, "Color space selection for self-organizing map based foreground detection in video sequences," in *2014 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2014, pp. 3347–3354.

[24] H. Sajid and S.-C. S. Cheung, "Universal multimode background subtraction," *Submitted to IEEE Transactions on Image Processing*, 2015.

[25] L. Maddalena and A. Petrosino, "A fuzzy spatial coherence-based approach to background/foreground separation for moving object

detection," *Neural Computing and Applications*, vol. 19, no. 2, pp. 179–186, 2010.

[26] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Static and moving object detection using flux tensor with split gaussian models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 414–418.

[27] M. Sedky, M. Moniri, and C. C. Chibelushi, "Spectral-360: A physics-based technique for change detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 399–402.

[28] K. Wang, C. Gou, and Y. Liu, "M4cd: A robust change detection method with multimodal background modeling and multi-view foreground learning," *Submitted to IEEE Transactions on Image Processing*, 2015.

[29] M. D. Gregorio and M. Giordano, "Wisardrp for change detection in video sequences," in *Submitted to the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[30] M. De Gregorio and M. Giordano, "Change detection with weightless neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 403–407.

[31] G. Ramírez-Alonso and M. I. Chacón-Murguía, "Auto-adaptive parallel som architecture with a modular analysis for dynamic object segmentation in videos," *Neurocomputing*, vol. 175, pp. 990–1000, 2016.

[32] G. Allebosch, D. Van Hamme, F. Deboeverie, P. Veelaert, and W. Philips, "C-efic: Color and edge based foreground background segmentation with interior classification," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2015, pp. 433–454.

[33] G. Allebosch, F. Deboeverie, P. Veelaert, and W. Philips, "Efic: edge based foreground background segmentation and interior classification for dynamic camera viewpoints," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2015, pp. 130–141.

[34] Y. Benezeth, P.-M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Comparative study of background subtraction algorithms," *Journal of Electronic Imaging*, vol. 19, no. 3, 2010.

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

[36] D. Liang and S. Kaneko, "Improvements and experiments of a compact statistical background model," *arXiv preprint arXiv:1405.6275*, 2014.

[37] S. Varadarajan, P. Miller, and H. Zhou, "Spatial mixture of gaussians for dynamic background modelling," in *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*. IEEE, 2013, pp. 63–68.

[38] B. Wang and P. Dudek, "A fast self-tuning background subtraction algorithm," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 395–398.

[39] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2. IEEE, 2004, pp. 28–31.

[40] X. Lu, "A multiscale spatio-temporal background model for motion detection," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 3268–3271.

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

[42] A. Miron and A. Badii, "Change detection based on graph cuts," in *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2015, pp. 273–276.

Worapan Kusakunniran received the B.Eng. degree in computer engineering from the University of New South Wales (UNSW), Sydney, Australia, in 2008, and the Ph.D. degree in computer science and engineering from UNSW, in cooperation with the Neville Roach Laboratory, National ICT Australia, Kensington, Australia, in 2013. He is currently a Lecturer with the Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom, Thailand. He is the author of several papers in top international conferences and journals. His current research interests include biometrics, pattern recognition, image processing, computer vision, multimedia, and machine learning.

Dr. Kusakunniran served as a Program Committee for the ICT International Student Project Conference (ISPC) 2016, the International Conference on Knowledge and Systems Engineering (KSE) 2015, the ACCV2014 Workshop on Human Gait and Action Analysis in the Wild: Challenges and Applications, the National Conference on Information Technology (NCIT) 2014, and the IEEE Workshop on the Applications of Computer Vision (WACV) 2013.

He has also served as a Reviewer for several international conferences and journals, such as the International Conference on Pattern Recognition (ICPR), the IEEE International Conference on Image Processing (ICIP), the IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), the Pattern Recognition (PR), the IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics (TSMCB), the IEEE Transactions on Image Processing (TIP), the IEEE Transactions on Information Forensics and Security (TIFS), the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), the Computer Vision and Image Understanding (CVIU), the Digital Signal Processing (DSP), the EURASIP Journal on Image and Video Processing (JIVP),

the Machine Vision and Applications (MVA), the International Journal of Automation and Computing (IJAC), and the IEEE Signal Processing Letters (SPL).

He was a recipient of ICPR Best Biometric Student Paper Award in 2010, ISPC 1^{st} Prize Award of Software Innovation Contest (Mobile Applications) in 2014, and ISPC 1^{st} Prize Award of Software Innovation Contest (ICT Applications) in 2016.

Rawitas Krungkaew received the B.Sc. degree in computer science from Mahidol University, Nakhon Pathom, Thailand, in 2009. He is currently working as a senior software engineer at Agoda Company Pte. Ltd. and also pursuing the M.Sc. degree in computer science, Faculty of Information and Communication Technology, Mahidol University. His current research interests include pattern recognition, image and video processing, and computer vision.