A Novel Motion Detection Method Using 3D Discrete Wavelet Transform

Yousefi, Sahar; Manzuri Shalmani, M.T. ; Lin, Jeremy; Staring, Marius

DOI
10.1109/TCSVT.2018.2885211

Publication date
2019

Document Version
Accepted author manuscript

Published in
IEEE Transactions on Circuits and Systems for Video Technology

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

A Novel Motion Detection Method Using 3D Discrete Wavelet Transform

Sahar Yousefi[1,2], M. T. Manzuri Shalmani[1], Jeremy Lin[3], and Marius Staring[2,4]

[1]Computer Engineering Department, Sharif University of Technology, Tehran, Iran
[2]a Division of Image Processing, Leiden University Medical Center, Leiden, The Netherlands
[3]PJM Interconnection, Audubon, USA
[4]Intelligent Systems Department, Delft University of Technology, Delft, The Netherlands

***Abstract*—The problem of motion detection has received considerable attention due to the explosive growth of its applications in video analysis and surveillance systems. While the previous approaches can produce good results, the accurate detection of motion remains a challenging task due to the difficulties raised by illumination variations, occlusion, camouflage, sudden motions appearing in burst, dynamic texture, and environmental changes such as those on weather conditions, sunlight changes during a day, etc. In this study, a novel per-pixel motion descriptor is proposed for motion detection in video sequences which outperforms the current methods in the literature particularly in severe scenarios. The proposed descriptor is based on two complementary three-dimensional discrete wavelet transforms (3D-DWT) and a three-dimensional wavelet leader. In this approach, a feature vector is extracted for each pixel by applying a novel three-dimensional wavelet-based motion descriptor. Then, the extracted features are clustered by the well-known K-means algorithm. The experimental results demonstrate the effectiveness of the proposed method compared to state-of-the-art approaches in several public benchmark datasets. The application of the proposed method and additional experimental results for several challenging datasets are available online.**

***Index Terms*—Motion detection; Dynamic texture; 3D-discrete Wavelet Transform; Wavelet leader**

I. Introduction

Motion detection in video sequences is the detection of moving objects throughout a subsequence of the frames. Over the past decade, this problem has attracted significant attention due to its wide range of applications in video surveillance, natural disaster investigation systems, and other areas. For this purpose, a wide variety of approaches have been proposed in the literature [1–12]. These approaches can be divided into: 1) spatial domain, and 2) frequency domain methods.

In spatial domain approaches, spatiotemporal descriptors are often used to model the motion by considering local motion patterns while ignoring holistic (global) motion patterns. St-Charles et al. [12] proposed SelfBalanced SENsitivity SEgmenter (SuBSENSE) as a pixel-level segmentation method that relies on spatiotemporal binary features and color information for change detection in video sequences. In their investigations, they used a spatiotemporal local binary similarity pattern (LBSP) for characterizing the pixel representations and then tuned the background parameters using pixel level feedback loops. In background tuning approaches, the background is modeled by a set of parameters which is updated by the history of recently observed pixel values. In these methods, the foreground detection depends on a decision threshold [13]. In [12], LBSP defines for each pixel p, a neighbor set $N(p)$ on each frame, and then assigns a binary pattern to the pixel p based on the differences of gray-levels between the neighboring pixels and p. If the illumination changes is non-uniform (i.e. the gray-level of a sub-set of neighbors changes), the binary pattern will be changed. Therefore, LBSP is not robust under non-uniform illumination changes. Moreover, in order to regularize the process and eliminate salt-and-pepper noise, SuBSENSE uses morphological operations and a median filter. The intrinsic noise attenuation property of these operations might inadvertently eliminate small moving objects. Furthermore, the background tuning process [12] is usually slow which makes adaptation to sudden illumination changes and burst motions difficult [9]. Bianco et al. [10] exploited genetic programming and combined state-of-the-art motion detection approaches to obtain the best solution. This method suffers from a heavy computational burden and there is no guarantee for finding the optimal solution.

On the contrary, frequency domain approaches can be considered holistic motion pattern extraction methods [14]. It has been reported that the two-dimensional discrete wavelet transform (2D-DWT) can be used for moving object detection [15, 16]. These methods are able to compare the 2D-DWT of the current frame with the previous frames and by using a threshold value to detect the motion. However, the intrinsic temporal dimension in the wavelet computation is not considered and the results are sensitive to the predefined threshold.

In another view, motion segmentation in video sequences can be divided into two categories: background-modeling and motion-modeling, where the former segments the moving regions in video sequences by comparing each new frame to a model of the background of the scene, while the latter addresses the motion segmentation problem with modeling the motion directly, i.e. without background modeling.

Although reasonable results can be obtained in the approaches mentioned earlier, accuracy in motion detection remains a challenging task due to the difficulties raised by illumination variations (such as those of weather conditions, sunlight changes during a day), occlusion, camouflage, sudden motion appearing in burst, dynamic texture, etc. Camouflage is a situation for which motion detection is difficult because

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

the color of foreground objects and that of the background are similar [17]. Spatiotemporal binary features can be combined with color information to detect the camouflaged foreground objects [12, 18]. Ramirez et al. [17] proposed a thresholding approach which tunes the values of the thresholds based on the analysis of the global Hue histogram of only the initial frame in order to identify if a color predominates on the scene. Due to considering one threshold value for all the frames based on only the initial frame, this method does not generalize to other frames or regions on the same frame with different intensity characteristics. A second challenge which is limitedly addressed in the literature [9] is the appearance of sudden motions which occurs in bursts. Moving escalators, swirling wheels, fans, swinging foliage are examples of burst motion. Background tuning approaches [9], where the background estimation is updated using the history of the values of the pixels over the time [14, 19–22], are sensitive to burst motion. Liang et al. generated a frequency and speed adaptive background model for these approaches [3, 9]. Dynamic texture detection is another issue which we considered in this work. Dynamic texture refer to every texture with motion in a sequence of video frames such as fire plume, water stream, smoke, etc. Dynamic texture detection has gained a great deal of attention recently [14, 19–21, 23]. To the best of our knowledge, none of the presented approaches however considered the issue of illumination changes for dynamic texture detection.

In this investigation, a novel motion-modeling method for detecting the motion in video sequences is proposed that defines a pixel-based feature descriptor based on the 3D-DWT and a recent development from the wavelet community, i.e. wavelet leaders, which we exploit for motion representation in video sequences. Wavelet leaders overcome the problem of a large number of close-to-zero wavelet coefficients [24]. Using the 3D-DWT rather than the 2D-DWT allows us to consider the motion continuously over time in the video frames. As the wavelet coefficient-based descriptors describe motion patterns based on decomposing the signal into the frequencies at multi-scales analysis, using a 3D-DWT can represent the spatial and temporal motion information together. This makes the developed approach applicable to dynamic texture detection and robust to sudden motion. Moreover, the proposed frequency-based features can provide a high degree of insensitivity to camouflaged foreground objects.

The three main contributions of our work are as follows:

1) We developed a spatial frequency per-pixel feature extraction method based on the high-pass/low-pass 3D-DWT accompanied by wavelet leaders for the first time in order to achieving proper motion detection results in video sequences. By using a 3D-DWT, continuity of timing information is provided by the proposed motion descriptors;
2) We developed a novel method that not only detects regular motion, as in previous works, but also sudden motion appearing in bursts. Also, the method can deal with dynamic textures;
3) We developed a robust and effective approach, which outperforms previous methods for videos taken under challenging conditions, such as varying weather conditions, illumination variations, camouflaged foreground objects, etc.

II. Background

The 2D-DWT is widely used for moving texture detection such as smoke detection [25], fire detection [26], etc. Demonceaux et al. [27] proposed the combination of a 2D-DWT and hierarchical Markov random fields for motion detection. In order to overcome the problem of temporal aliasing, an estimation of dominant motion on several image resolutions is obtained. The results however exhibited obvious jaggedness of the boundaries. In our work, by using 3D wavelet-based descriptors, the problem of the jagged boundaries has been solved. In this section an overview of the 3D-DWT as a separable filter and wavelet leaders is provided.

A. Three dimensional discrete wavelet transform

The 3D-DWT has been employed for different purposes. In [28] a 3D wavelet transform has been used for scene change detection. However, different from the proposed approach, it is used for classifying the motion of entire frames into three broad classes: frames with motion, frames with a gradual transition, and static frames. In this paper, we detect the specific area where motion is observed. In [29], unlike our work, uses a 3D-DWT for video coding not for motion detection. In fact, they used a figure-background decision based on a long-term memory of static pixels that also takes the motion information into account for processing the foreground and background and estimating the motion. Then, a 3D wavelet transform is used for coding the residual signal for reconstructing the video sequences. Our method has employed the 3D wavelet transform and wavelet leader concept in order to represent motion in temporal and spatial dimensions for object motion detection in video frames, and this has not been addressed before in this way.

The 3D-DWT can be represented as a separable filter, thus reducing the cost of computation. The separable 3D-DWT filters can be written as the product of three simpler 1D filters: two 1D spatial and one 1D temporal DWT filter [30]. In the 1D wavelet transform, a function $f(t)$ is analyzed by:

$$f(t) = \sum_{\tau,s} \Psi_t(\tau, s)\varphi_t(\tau, s), \tag{1}$$

where $\Psi_t(\tau, s)$s are the wavelet coefficients estimated by:

$$\Psi_t(\tau, s) = \int_{-\infty}^{\infty} f(x)\varphi_t(\tau, s)dt. \tag{2}$$

Each of the wavelet coefficients $\Psi_t(\tau, s)$ represents the resemblance of the function $f(t)$ to the wavelet bases $\varphi_t(\tau, s)$ at a specific translation τ and scale s.

In this paper, we use Coiflet-like [31] nearly symmetric orthogonal wavelet bases with a magnitude and group delay flatness specification, which has been proposed by Abdelnour et al. [32]. Using the multi-scale wavelet decomposition scheme, a hierarchy of localized sub-functions at different spatial frequencies can be found. Suppose we have a volume

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

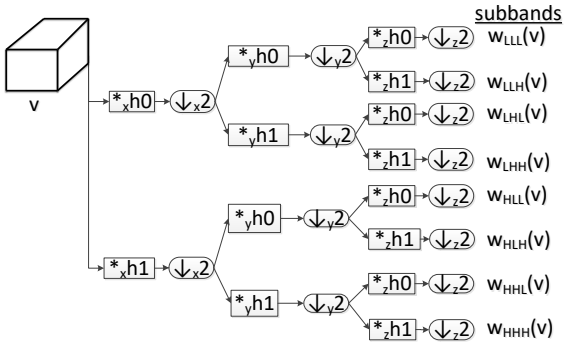

Fig. 1: Wavelet decomposition using the separable 3D-DWT filter. First the volume V is decomposed in a low-frequency channel L and a high-frequency channel H, by repeating this process for each of sub-volumes four different sub-volumes (LL, LH, HL, HH) are computed, and by repetition we obtain eight subbands (LLL, LLH, LHL, LHH, HLL, HLH, HHL, HHH).

V, a 3D-DWT decomposes the volume V into one low-frequency channel w^1_{LLL}, seven strict high-frequency channels $w^S_{o,l}$, and multiple non-strict high-frequency channels $w^s_{o,l}$ for $s \in \{1,\dots,S-1\}$, where s is from the scale set, S is the coarsest scale, $o \in O$ is from the orientation set, where $O = \{vertical, horizontal, diagonal\}$, and $l \in \Gamma$ is from the level set, where $\Gamma = \{up, down\}$. Decomposing a volume V into the wavelet coefficient uses a recursive process function $\Phi^S(V)$, in which for each inter-volume v and each scale $s \in \{1,...,S\}$, $\Phi^s(v)$ is defined as:

$$\Phi^s(v) = \begin{cases} \psi^s(v), s = 1 \\ \{\psi^s(v) - w^s_{LLL}(v)\} \cup \Phi^{(s-1)}\left(w^s_{LLL}(v)\right), s > 1 \end{cases} \quad (3)$$

and ψ^s for each inter-volume v is equal to:

$$\psi^s(v) = \{w^s_{LLL}(v), w^s_{LLH}(v), w^s_{LHL}(v), w^s_{LHH}(v), w^s_{HLL}(v), w^s_{HLH}(v), w^s_{HHL}(v), w^s_{HHH}(v)\}. \quad (4)$$

Regarding filter concepts, Fig. 1 illustrates the 3D-DWT filter banks for $\Phi^{(S=1)}(V)$. The figure indicates the sub-volumes of each filter for one scale and a volume V.

B. Wavelet leader

In order to improve the robustness of the descriptors of the wavelet coefficients, we use wavelet leaders [24]. Wavelet leaders are another wavelet-based measurements which were defined by Jaffard et al. for the first time [24]. In the literature, wavelet leaders are used for various applications [14, 33–36]. Wavelet leaders are defined as the maximum magnitude of the wavelet coefficients in a local spatial neighborhood through all scales. As mentioned before, wavelet leaders obviate the problem of many close-to-zero wavelet coefficients. This makes the motion feature descriptors more robust. In this paper, we propose three-dimensional wavelet leader pyramids. The wavelet leader for a pixel p which is the center of a cubic neighbourhood Λ_p and scale $1 \le s \le S$ is defined as:

$$w^s_{leader}(p) = \max_{o\in O} \max_{l\in L} \max_{r'\in\Lambda_p} |w^s_{o,l}(r')|. \quad (5)$$

III. The Proposed Method

In this section, the proposed approach is provided. A flow diagram of the method is shown in Fig. 2. As illustrated, the input is a sequence of video frames, $\{I_t | t \in [1,T]\}$ in which t is a temporal variable, and the output is a sequence of the label fields, $\{\ell_t | t \in [1,T]\}$.

A. Cubic patch extraction

As motion can be done within spatial and temporal dimensions in videos, motion descriptors must be defined for both domains. Hence, the process of 3D-DWT feature extraction in our paper is patch-based. As illustrated in Fig. 2 the first step of the flow diagram is patch extraction. In this step, x and y depict the spatial domain and t depicts temporal domain. Therefore, in addition to considering temporal coherency, descriptors capture both spatial and temporal motion patterns. This process is done for each pixel p on the frames, by defining a cubic neighbourhood set Λ_p, where p is the central pixel. Fig. 3 represents Λ_p of size $4\times4\times4$. In this figure, the volume is defined by selecting a 4×4 neighborhood on a sequence of frames of length 4. In Section IV-C, the evaluation of the proposed method for various patch sizes will be investigated. The results demonstrate that the patch size of $4 \times 4 \times 4$ outperforms other patch sizes in accuracy.

B. 3D-Discrete Wavelet Transform (3D-DWT)

After extracting the patches, we use 3D-DWT. Applying 3D-DWT on a three-dimensional volume computes the approximation coefficient and seven detail coefficients in seven different directions. These coefficients in different directions describe the motion within the spatial domain and appearance of the video sequences across time. As mentioned before, by considering a separable 3D-DWT the process of computing the wavelet coefficients can be decoupled into two 1D spatial DWTs and one 1D temporal DWT. The spatial DWTs consider the holistic motion patterns in space while the temporal DWT considers the holistic motion pattern across time. In the proposed method, the 3D-DWT is applied to the cubic patches, which is illustrated in Fig. 2, Step 2. Then, for each scale the wavelet leaders are computed. In this step, the original patch is high-pass filtered, yielding the seven larger volumes labeled in green, each describing local changes in details in the original volume. It is then low-pass filtered and down scaled, yielding an approximation volume; this volume is high-pass filtered to produce the seven smaller detail volumes labeled in blue, and low-pass filtered to produce the second approximation volume. This iterative process is repeated for the second approximation volume to produce seven smaller detail volumes labeled in red. Then the second approximation volume is low-pass filtered to produce the final approximation volume in the upper-left. In this step, the wavelet leader for each scale is computed using Equation (5). For this example, the yellow, orange and purple patch depict the 3^{rd}-scale, 2^{nd}-scale, and 1^{st}-scale wavelet leaders, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

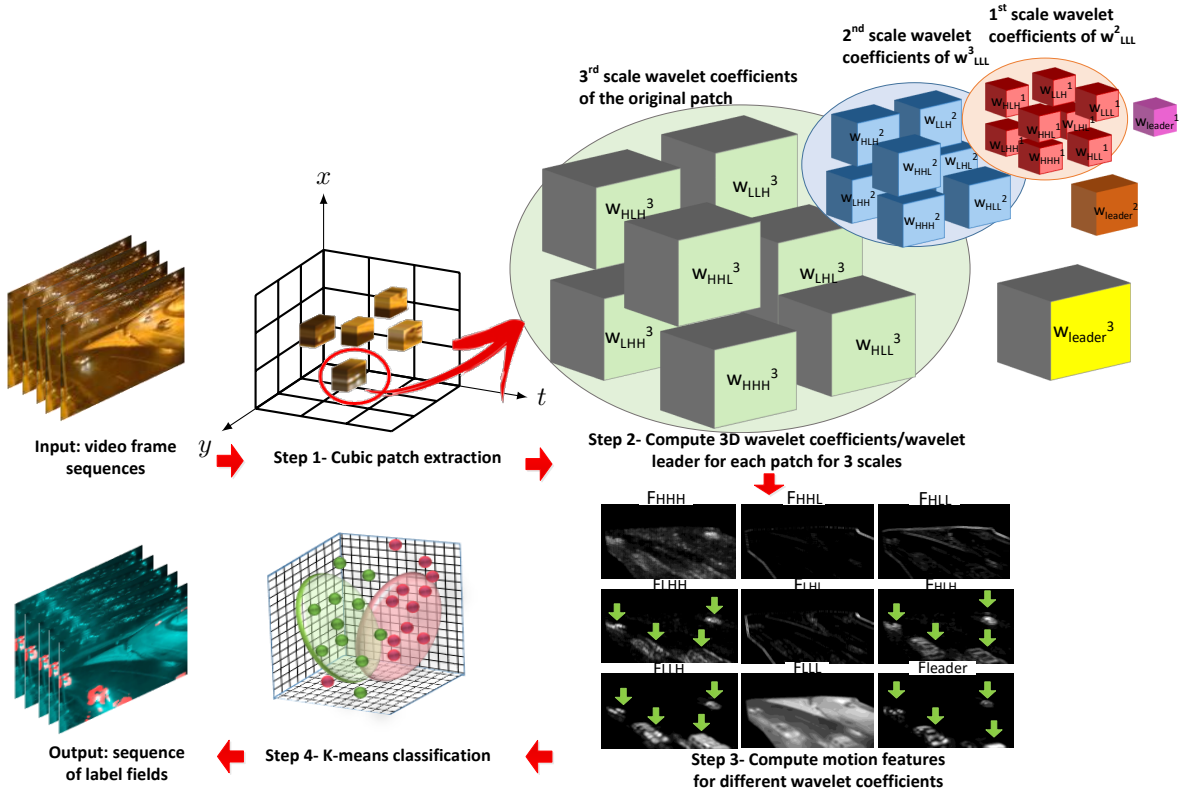

Fig. 2: Flow diagram of the proposed method. The input is a sequence of video frames; Step 1: Extracting the cubic patches; Step 2: Applying the 3D-DWT on each patch for computing wavelet coefficients, and wavelet leaders. The seven green patches are the wavelet coefficients extracted from the original patch, the yellow patch is the wavelet leader of 3^{rd} scale, the seven blue patches are the wavelet coefficients extracted from W^3_{LLL}, the orange patch is the wavelet leader of 2^{nd} scale, the eight cubic red patches are the wavelet coefficients extracted from W^2_{LLL}, and the purple patch is the wavelet leader of 1^{st} patch. Step 3: Computing the feature descriptors using the wavelet coefficients and the wavelet leaders, here the value of features for all the wavelet coefficients and the wavelet leaders are shown. The green arrows depict the feature descriptor values for the moving objects in the video sequences. As can be seen, the feature descriptors $F_{LHH}, F_{HLH}, F_{LLH}, F_{leader}$ can model the motion pattern obviously; Step 4: Applying *K-means* classification. Finally, the result of the classifier which illustrate moving objects (red regions) and static regions (green regions).

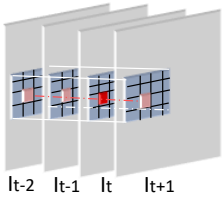

Fig. 3: Neighbourhood Λ_p of size $4 \times 4 \times 4$ on a sequence of lattices of the subsequent frames, the red pixel shows p and the other pixels on the blue grids demonstrate the neighbourhood set Λ_p. I_t is the frame at time t which contains p.

C. Feature Extraction

In order to obtain proper results and prevent jagged boundaries, the feature vectors are described by a pixel-based process over the wavelet coefficients obtained in the previous step. Zhang et al. proposed some feature vectors for the 2D-DWT [37]. In our work, we propose feature vectors based on the 3D-DWT decomposition coefficients and a 3D-wavelet leader. By using 3D-wavelet coefficients, we can model the holistic motion patterns in different directions. By applying 3D-wavelet transform on a volume or a cubic patch, the approximation coefficient and seven detail coefficients in seven different directions in a 3D space are computed. The high-pass wavelet coefficients can model the motion patterns over a sequence of frames. We define the descriptor vectors as:

$$F_{\hbar}(p) = \frac{1}{\sum_{s=1}^{S} s} \sum_{s=1}^{S} s \sqrt{\sum_{r' \in \Lambda_p} \left(w^s_{\hbar_{r'}}(r^s) \right)^2}, \qquad (6)$$

where p is the central pixel on a neighborhood cube Λ_p, r' the neighbors of p, $\hbar \in \Delta$ where

$$\Delta = \{LLL, LLH, LHL, LHH, HLL, HLH, HHL, HHH, leader\} \qquad (7)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

and r^s is defined as:

$$r^s = \begin{cases} \Lambda_p, s = S \\ w_{\hbar}^{s+1}\left(r^{s+1}\right), 1 \leq s < S, \end{cases} \quad (8)$$

in which S is the coarsest scale, $w_{\hbar}^s$ demonstrates the wavelet sub-bands of the s^{th} scale.

In Equation (6), $\sqrt{\sum_{r' \in \Lambda_p}\left(w_{\hbar_{r'}}^s\left(r^s\right)\right)^2}$ computes the Euclidean norm on a $n-$space [38] for the s^{th} scale level and the $\hbar^{th}$ wavelet sub-bands, which n is the number of the neighbours of p. This distance gives the ordinary distance from an origin to the patch vector. This distance can be considered as the difference between the value of the central pixel and its neighbours in a patch which implies motion in a patch. Also, averaging over multiple scale levels in Equation (6) leads to apply the more significant wavelet components in finer scales. which consequently causes noise reduction. In this paper, in order to compute the feature descriptors, we use S scales for volumes of size $2^S \times 2^S \times 2^S$. In Fig. 2- Step 3 illustrates the feature vectors extracted from the wavelet coefficients using Equation (3) and the wavelet leaders using Equation (5) for a set of consecutive frames, using cubic patches of size $4 \times 4 \times 4$ and $S = 3$. As it is shown by the green arrows, the feature vectors extracted from high-pass wavelet coefficients contain w_{LHH}, w_{HLH}, w_{LLH}, and w_{leader}, illustrate motion patterns significantly. Hence, we use these feature vectors to model the motion patterns in three-dimensional space.

D. Classification

Ultimately, after feature extraction, the feature vectors are classified into two classes using K-means clustering: motion & zero-motion. In Fig. 2 the outputs are illustrated as a sequence of label fields, red regions demonstrate the moving objects and green regions demonstrate the static regions.

Fig. 2 explains the proposed method as follow. In the first step, the cubic patches are extracted; then the 3D-DWT of three scales is applied on each patch in order to compute wavelet coefficients using Equation (3), and wavelet leaders using Equation (5). Generally, for S scales $7 \times S + 1$ patches for the wavelet coefficients and S patches for the wavelet leaders are computed. Afterward, the feature descriptors based on Equation (6) using the wavelet coefficients and the wavelet leaders are computed. Then *K-means* classification is applied to classify motion vs. non-motion regions. Finally, the result of the classifier is a sequence of label fields which illustrate moving and static regions.

IV. Experimental results

In order to evaluate the performance of the proposed approach, this section reports the qualitative and quantitative results on video datasets including a variety environments. Also, the implementation of the proposed method in Matlab 8.3 and C, named the 3D-DWT motion detector (3D-DWT-MD) tool, is freely available at: https://github.com/yousefis/MotionDetector.

The qualitative and quantitative results are reported by comparing them with the results of several unsupervised approaches including background-modeling approaches: SOBS-CF [2], CP3-online [9], SUBSENSE [12], C-EFIC [39], GMM|Zivkovic[40], AMBER [41], and motion-modeling approaches: CwisarDH [4], IUTIS-3 [10], 2D-DWT [15], AAPSA [17]. We optimized a hyper-parameter (the patch size) on an independent dataset (LASIESTA), and evaluated the proposed method on three other datasets:

1) For finding the optimal patch size we used LASIESTA dataset (available at https://www.gti.ssr.upm.es/data/lasiesta_database.html), which is composed by many real indoor and outdoor sequences organized in different categories [42]. Since we do not consider camera motion, we used the sequences which were captured by a static camera. The indoor part includes *Simple sequences*, *Camouflage*, *Occlusions*, *Illumination changes*, *Modified background*, *Bootstrap*, and the outdoor part contains *Cloudy conditions*, *Rainy conditions*, *Sunny conditions*.
2) In order to evaluate our method for motion detection under various light conditions, we use the CD.net2014 dataset (available at www.changedetection.net). This dataset provides a realistic, camera-captured, diverse set of videos which contains several video categories with 4 to 6 video sequences in each category [43]. Furthermore, each category is accompanied by accurate ground truth segmentation and annotation of change or motion areas for each video frame. In order to compare our method with the previous methods, we obtained the results of the previous methods which were reported and maintained on http://dspl.ce.sharif.edu/motiondetector.html. For this goal, we compare the proposed method with the previously reported unsupervised methods which are published in first-tier conferences and journals in recent years. Since we do not consider camera motion, we report experimental results for the video sequences which are captured by a static camera. For this purpose, we examine the proposed method for four different categories include: *NightVideos*, *Thermal*, *IntermittentObjectMotion* and *Baseline*, which contain nineteen video sequences. The video sequences are *highway*, *Parking*, *StreetLight*, *AbandonedBox*, *WinterDriveway*, *TramStop*, *sofa*, *Park*, *LakeSide*, *Corridor*, *diningRoom*, *Library*, *BridgeEntry*, *busyBoulvard*, *FluidHighway*, *StreetCornerAtNight*, *TramStation*, *WinterStreet*.
3) In order to examine the proposed method for dynamic texture detection, the Dyntex dataset [20] is used. Dyntex is a comprehensive database of dynamic textures providing a large and diverse database of high-quality dynamic textures, which have been de-interlaced with a spatiotemporal median filter [20]. The dynamic texture sequences have been acquired using a SONY 3 CCD camera using a tripod. The sequences are recorded in PAL format (720×576). In this work, we evaluate the proposed method for ten different

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

video series of this dataset contain: moving smoke ($648ea10$, $649hb10$, $73v192u$, and $57db110$), washing machine ($64ca510$), CD driver ($64bac10$), waving water ($6ame200$ and 6482420), shower ($56ub110$), and moving flame ($64cad10$).

4) The UCSD pedestrian dataset [19] is another dataset which is commonly used for motion detection evaluation. The dataset contains video of pedestrians on UCSD walkways, taken from a stationary camera with two different viewpoints. In this paper, we compare the results of the proposed method to the results of MDT [19] (available at http://visal.cs.cityu.edu.hk/).

A. Qualitative Comparison

In this section, a qualitative comparison of the proposed method with various methods is provided. From the viewpoint of occlusion, Fig. 4 illustrates the comparison of the moving object segmentation between the proposed method and the previous approaches, including CP3-online [9], IUTIS-3 [10], and SUBSENSE [12], for two frames of the *WinterStreet* sequence. In the figure, the red segments are the ground truth masks, the blue segments illustrate the moving regions, the green segments represent the static regions, and the yellow circles highlight the differences between the segmentation results. As can be seen, the mentioned methods consider two consequent moving objects as one object while 3D-DWT-MD can distinguish the individual moving objects.

Fig. 5 illustrates more qualitative comparisons of our method with the aforementioned approaches, for various frames of *WinterStreet* sequence of the CD.net 2014 dataset. As shown in the results, despite severe environmental conditions raised by video acquisition and car light at night, unlike the other methods the proposed method can deal with the occlusion properly. Another qualitative comparison, for various frames of *StreetCornerAtNight* sequence of the CD.net 2014 dataset, is indicated in Fig. 6. The results of the proposed method are compared with CP3-online [9] and SUBSENSE [12]. As shown in the results, the proposed method can overcome the motion detection problem at night light properly. Moreover, Fig. 7 illustrates another qualitative comparison of the proposed method with CP3-online [9], SUBSENSE [12], and AAPSA [17] for various frames of the busyBoulvard sequence of the CD.net 2014 dataset. As these results indicate, the proposed method can overcome the illumination variations, occlusion more robustly.

Fig. 8 illustrates a qualitative comparison of the proposed method with CP3-online [9], AAPSA [17] and 2D-DWT [15] for various frames of *Corridor* and *Park* sequence in CD.net 2014 dataset. The results indicate the appropriate ability of the proposed method in difficulties raised by camouflage.

Fig. 9 illustrates the motion detection results of CP3-online [9], SUBSENSE [12], and our method respectively for various frames of the *blizzard*, *streatlight*, and *Parking* sequences in the CD.net 2014 dataset. As the results indicate, while the other mentioned methods can not detect moving objects in these video scenarios, the proposed method can detect tiny moving objects perfectly. Furthermore, Fig. 10 indicates a qualitative comparison of the motion detection results with CP3-online [9], for various frames of *busStation* sequence of the CD.net 2014 dataset. As can be seen, the results of 3D-DWT-MD is much more robust. Fig. 11 indicates another qualitative comparison of the motion detection results with MDT [19] for some frames of the sequences of the UCSD pedestrian dataset. In this figure, the green segments represent moving object and the yellow segments represent static regions. Results indicate that the efficiency is improved by the proposed method. Finally, Fig. 12 illustrates the motion detection results of the proposed method on various frames of sequences from the Dyntex dataset. The sequences contain dynamic textures like smoke, fire, and waving water. Also, burst motions contain disk driver and washing machine. The results indicate that the proposed method can be used for dynamic texture segmentation in video sequences.

B. Quantitative measurements

For quantitative comparison, various evaluation metrics contain Recall (Re); Specificity (Sp); False Positive Rate (FPR); False Negative Rate (FNR); Percentage of Wrong Classifications (PWC); F-measure, and Precision will be used. Recall can be seen as the completeness of the moving object. Specificity can be seen as the completeness of background. FPR is the rate of the background which is detected as the foreground incorrectly, and FNR, the rate of the foreground which is detected as the background incorrectly. The PWC measurement is the percentage of the foreground and background which is detected incorrectly. Finally, the F-measure is a weighted harmonic mean of the Precision and Recall. For the aforementioned measures, zero is the best value for FPR, FNR, and PWC, while one is the best value for Recall, Specificity, and F-measure.

C. Quantitative comparison

The most important parameter of the model is the cubic patch size for computing the wavelet coefficients. For tuning this parameter, we evaluate the proposed algorithms on the LASIESTA database for the different cubic patch sizes given in Table I. In this table the scale decomposition levels are shown. The patch sizes are defined such that they are a power of 2 in spatial and temporal dimensions. The number of scales is defined by the logarithm of the patch size base 2. In this section the different patch sizes are used for finding the optimal patch size, on this independent dataset.

Table II illustrates the average of the different measures containing Re, SP, FPR, FNR, PWC, Precision, F-measure of the proposed approach with different scales

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

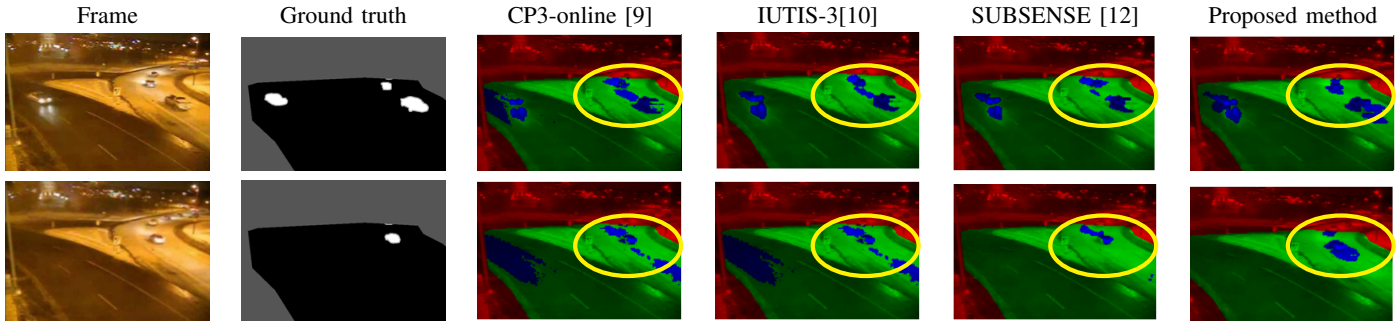

Fig. 4: Qualitative comparison of the various approaches in the presence of occlusion. Shown are two frames (#990 & #1027) from the CD.net 2014 dataset, *WinterStreet* sequence.

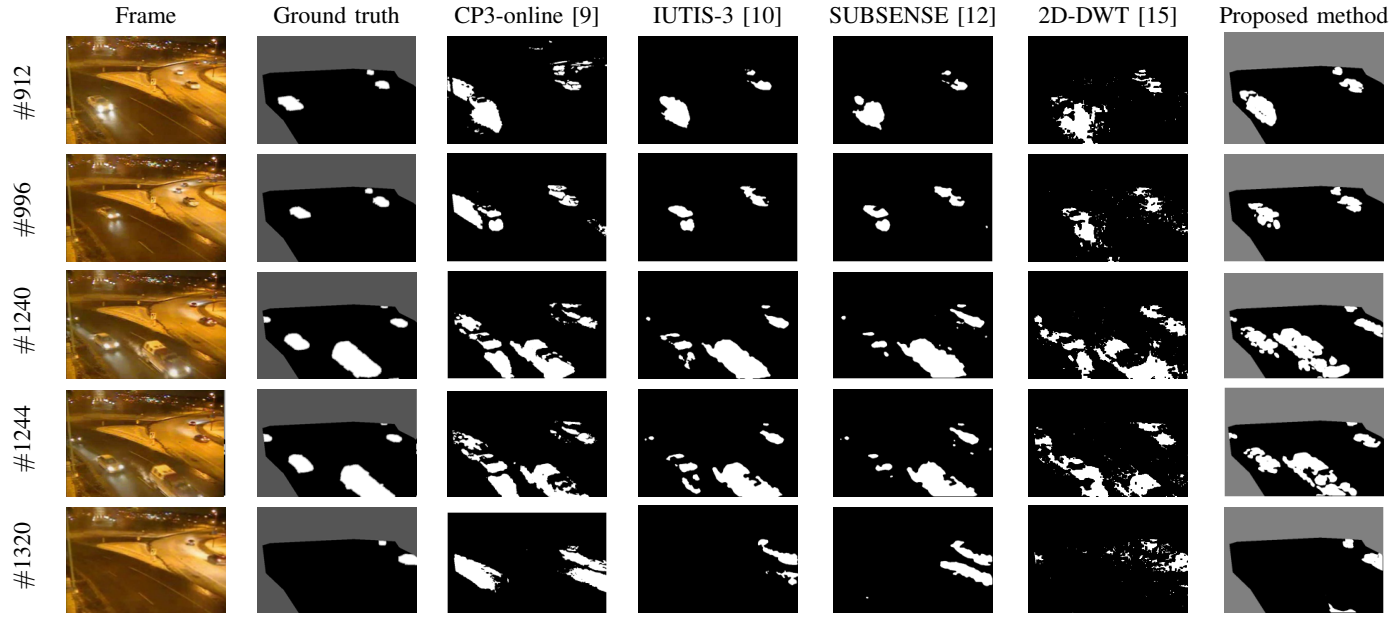

Fig. 5: Qualitative comparison of the various approaches for several frames of the *WinterStreet* sequence of the CD.net 2014 dataset. The gray regions indicate the ground truth masks. These masked for the results of the other methods are applied as zero masks. More comparisons without applying the masks are available at http://dspl.ce.sharif.edu/motiondetector.html.

TABLE I: The sizes of the cubic patches and their decomposition scales which are used for the experimental results.

Patch sizes	Scale	Decomposition levels
$2 \times 2 \times 2$	1	$1 \times 1 \times 1$
$4 \times 2 \times 2$	1	$2 \times 1 \times 1$
$8 \times 2 \times 2$	1	$4 \times 1 \times 1$
$2 \times 4 \times 4$	1	$1 \times 2 \times 2$
$4 \times 4 \times 4$	2	$2 \times 2 \times 2 \rightarrow 1 \times 1 \times 1$
$8 \times 4 \times 4$	2	$4 \times 2 \times 2 \rightarrow 2 \times 1 \times 1$
$2 \times 8 \times 8$	1	$1 \times 4 \times 4$
$4 \times 8 \times 8$	2	$2 \times 4 \times 4 \rightarrow 1 \times 2 \times 2$
$8 \times 8 \times 8$	3	$4 \times 4 \times 4 \rightarrow 2 \times 2 \times 2 \rightarrow 1 \times 1 \times 1$

and sizes, for the LASIESTA dataset. Also, Fig. 13 shows a qualitative example of the comparison of the results' quality depending on the patch size. Results indicate that a patch size of $4 \times 4 \times 4$ produces the best F-measure value.

Table III gives a quantitative comparison with various approaches for frames of the mentioned eighteen different video sequences of the CD.net 2014 dataset. In this experiment, the cubic patch sizes are $4 \times 4 \times 4$ and the decomposition scale is 2. As the results show, the average and standard deviation values of the measures, containing Re, Sp, FNR, PWC, Precision and

TABLE II: The average of the different measures and computation time at different scales and patch sizes for the LASIESTA dataset. The best values are shown in bold. More results are available at http://dspl.ce.sharif.edu/motiondetector.html.

Patch Size	Re	SP	FPR	FNR	PWC	Precision	F-measure
$2 \times 2 \times 2$	0.43	**0.99**	**0.01**	0.57	3.61	**0.70**	0.53
$4 \times 2 \times 2$	0.52	0.99	0.01	0.48	3.40	0.69	0.59
$8 \times 2 \times 2$	0.66	0.98	0.02	0.34	3.58	0.61	0.64
$2 \times 4 \times 4$	0.59	0.99	0.01	0.41	3.26	0.68	0.63
$4 \times 4 \times 4$	0.74	0.98	0.02	**0.26**	**2.84**	0.69	**0.71**
$8 \times 4 \times 4$	0.67	0.98	0.02	0.33	3.56	0.61	0.64
$2 \times 8 \times 8$	0.69	0.98	0.02	0.31	3.73	0.59	0.64
$4 \times 8 \times 8$	0.70	0.97	0.03	0.30	3.99	0.57	0.63
$8 \times 8 \times 8$	**0.75**	0.96	0.04	0.25	5.01	0.48	0.59

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

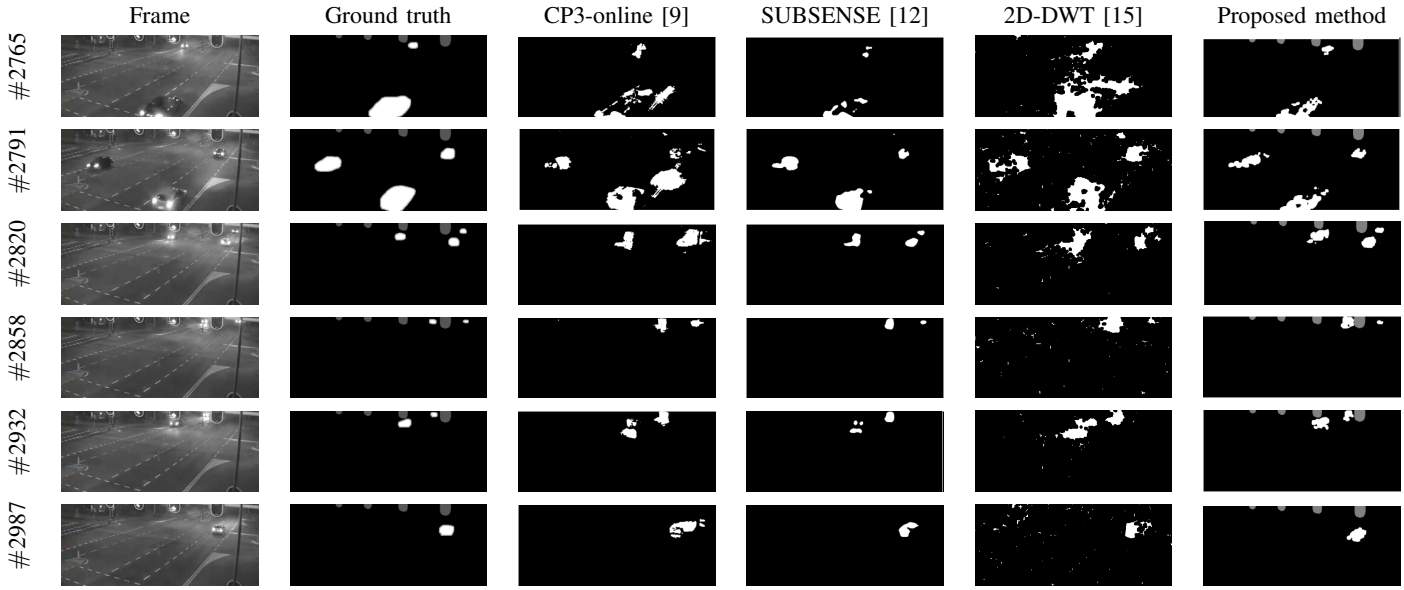

Fig. 6: Qualitative comparison of the results with different approaches for *StreetCornerAtNight* sequence of the CD.net 2014 dataset. The gray regions indicate the ground truth masks. These masked for the results of the other methods are applied as zero masks. More comparisons without applying the masks are available at http://dspl.ce.sharif.edu/motiondetector.html.

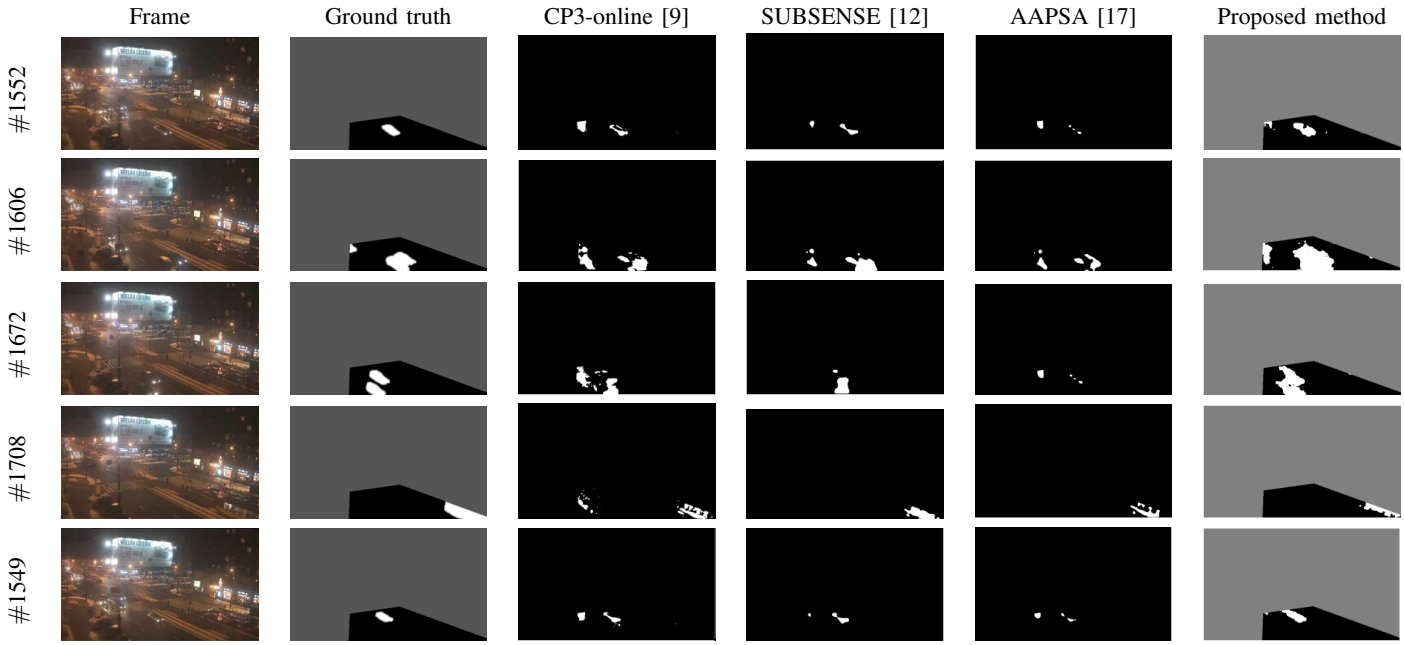

Fig. 7: Qualitative comparison of the results with different approaches for *BusyBoulvard* sequence of the CD.net 2014 dataset. The gray regions indicate the ground truth masks. These masked for the results of the other methods are applied as zero masks. More comparisons without applying the masks are available at http://dspl.ce.sharif.edu/motiondetector.html.

F-measure of the proposed method, for four different mentioned categories, are equal to 0.82, 0.94, 0.06, 0.17, 4.11, 0.79 and 0.78 respectively. According to the value of F-measures, these quantitative results indicate an substantial improvement compared with the previous approaches.

Fig. 14 illustrates a comprehensive quantitative comparison of the average of the different measurements between the proposed method and a substantial number of existing unsupervised approaches for video sequences of CD.net 2014. As it can be comprehended, the F-measure for the proposed method with 0.81 has the highest median value. Also, it tends to to have the most narrow fences in comparison with the other approaches which shows that the proposed method outperforms the other methods remarkably. Moreover, it is clear that the fences defined by the Presicion and Recall are far too small in comparison to the previous methods. It is clear

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

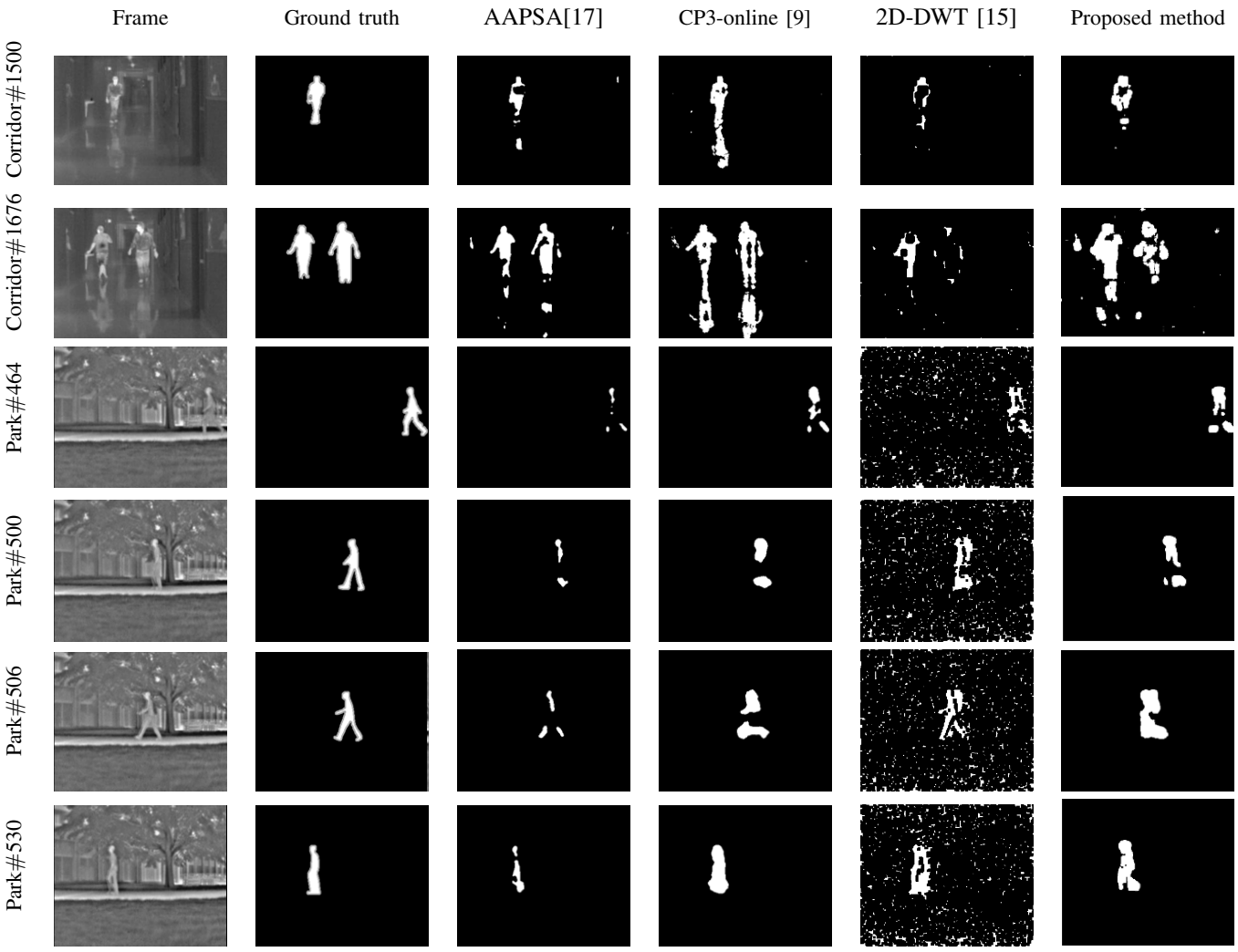

Fig. 8: Qualitative comparison of the results with different approaches *Thermal* sequence of CD.net 2014 dataset

TABLE III: Comparison of the measures of different approaches for eighteen video sequences belong to four categories of CD.net 2014 dataset (μ stands for mean and σ stands for standard deviation)

. The best results are shown in bold.

Method	Re		SP		FPR		FNR		PWC		Precision		F-measure	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
SOBS-CF [2]	0.71	0.22	0.96	0.05	0.04	0.05	0.29	0.22	4.96	4.40	0.61	0.31	0.59	0.26
CwisarDH [4]	0.57	0.28	0.99	0.01	**0.01**	**0.01**	0.43	0.28	3.38	3.65	0.71	0.25	0.59	0.27
CP3-Online [9]	0.74	0.19	0.93	0.15	0.07	0.15	0.26	0.19	6.99	1.25	0.56	0.27	0.60	0.24
IUTIS-3 [10]	0.69	0.22	0.99	0.01	0.01	0.01	0.31	0.22	**2.70**	**2.73**	0.74	0.25	0.68	0.22
SuBSENSE [12]	0.72	0.20	0.99	0.01	0.01	0.01	0.28	0.20	3.13	3.04	0.73	0.24	0.69	0.21
AAPSA [17]	0.51	0.25	**0.99**	**0.01**	0.01	0.01	0.49	0.25	3.50	3.17	0.70	0.27	0.55	0.25
C-EFIC [39]	0.79	0.15	0.97	0.08	0.03	0.08	0.21	0.15	4.11	7.01	0.71	0.25	0.71	0.21
GMM|Zivkovic [40]	0.55	0.20	0.98	0.03	0.02	0.03	0.45	0.19	4.66	4.44	0.65	0.28	0.54	0.21
AMBER [41]	0.70	0.23	0.97	0.04	0.03	0.04	0.30	0.23	4.03	4.48	0.67	0.30	0.63	0.26
2D-DWT [15]	0.39	0.19	0.98	0.02	0.47	0.33	0.61	0.19	4.82	5.40	0.49	0.30	0.36	0.13
Proposed method	**0.82**	**0.14**	0.94	0.21	0.06	0.21	**0.18**	**0.14**	4.11	9.39	**0.79**	**0.20**	**0.78**	**0.13**

that for the proposed method the FNR median has the lowest value and its fences are more narrow than most of the other methods. Despite the improvements in the mentioned measurements, the value of Specificity and PWC are equal to the other methods. Therefore this figure indicates that the proposed method improves on the other approaches significantly.

V. Discussion

In this work, a robust motion detection method in video sequences has been presented. Its core is a 3D-DWT based feature descriptor which is resistant to illumination changes. By using separable filter banks the DWT can be directly applied to one-dimensional subbands. It is well known that 1D wavelet transformations for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

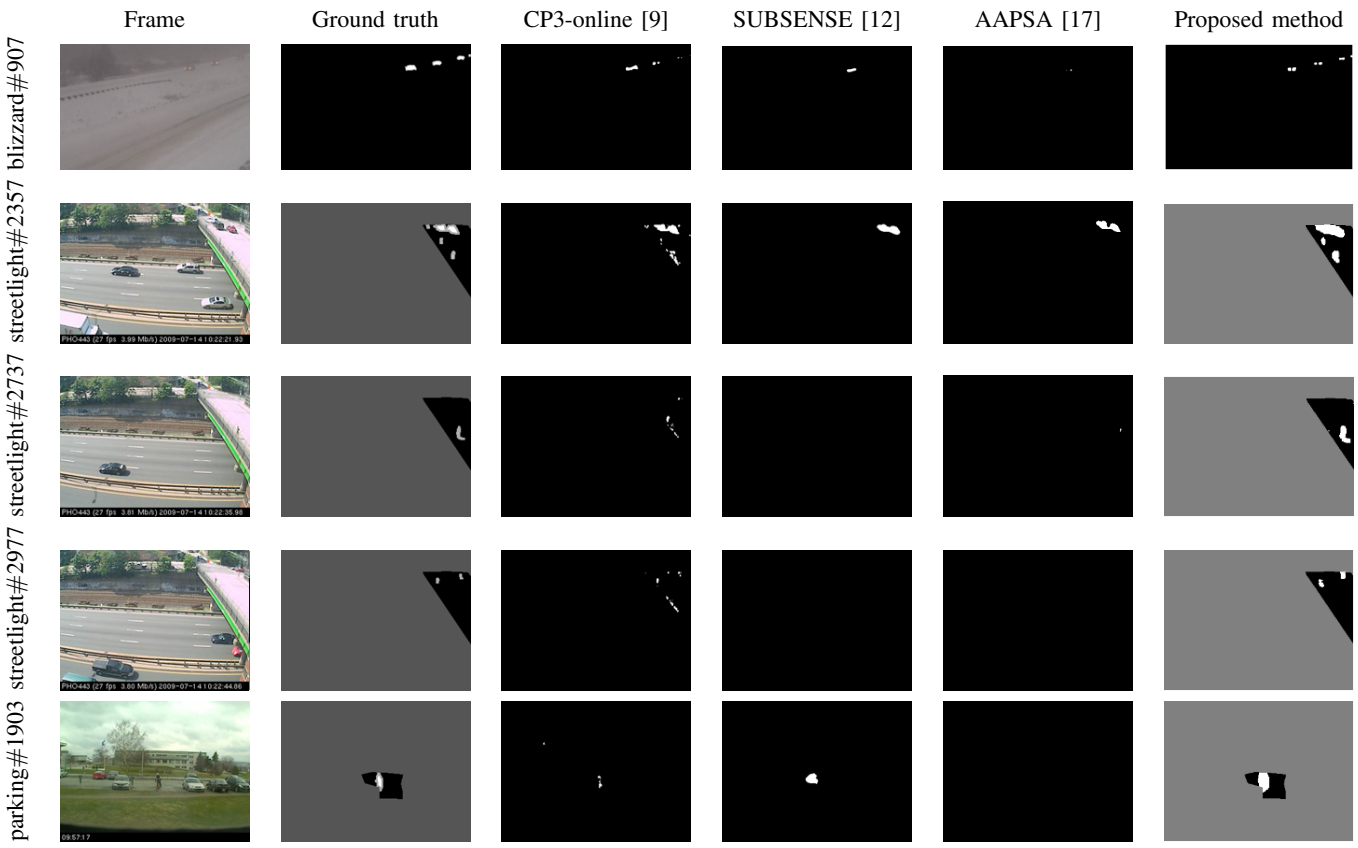

Fig. 9: Qualitative comparison of the segmentation results with different approaches for various frames of *IntermittentObject-Motion* sequence of the CD.net 2014 dataset, as it can be perceived, the competitive methods do not detect moving objects while the proposed method even can detect tiny moving objects properly. The gray regions indicate the ground truth masks. These masked for the results of the other methods are applied as zero masks. More comparisons without applying the masks are available at http://dspl.ce.sharif.edu/motiondetector.html.

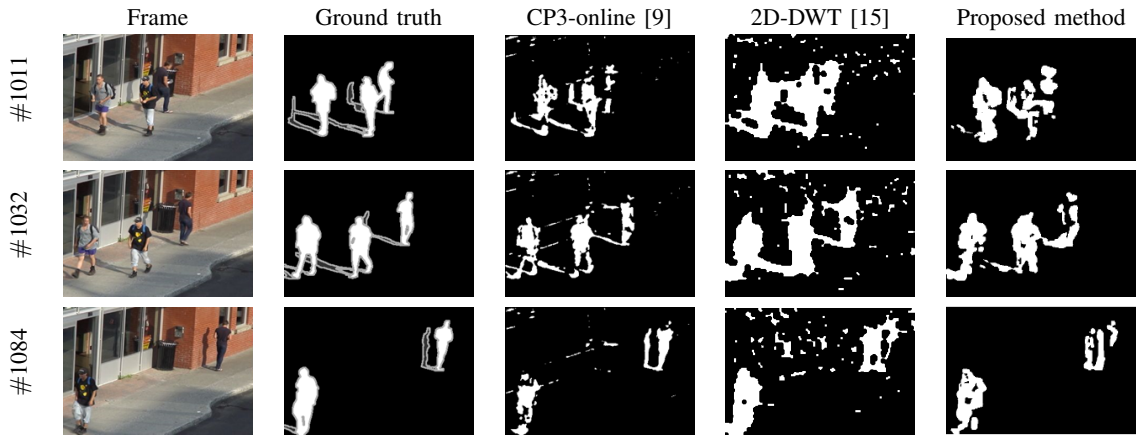

Fig. 10: Qualitative comparison of the results with different approaches *BusStation* sequence of the CD.net 2014 dataset

a signal with n points can be accomplished in $\mathcal{O}(n)$ time [44]. As we used a separable DWT, the time complexity of the proposed method can be computed linearly with respect to the patch size for each patch. For a patch of size $n = p_x \times p_y \times p_t$ and for S scales the complexity of calculating the wavelet coefficients is equal to $T(n, S) = n \sum_{s=0}^{S-1} \frac{1}{2^{3s}}$. Hence, the time complexity of the proposed method for p_t frames with size $N \times M$ is equal to $T_{3D}^{DWT} = \mathcal{O}(T(n, S) \times N \times M)$, in which n and S are not large numbers ($\ll N, M$, w.r.t Table I). Hence, T_{3D}^{DWT} can be simplified to $\mathcal{O}(N \times M)$, which is equal to the time complexity of the 2D-DWT. In the remainder of this section first the limitations of the proposed method are discussed, then future works

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

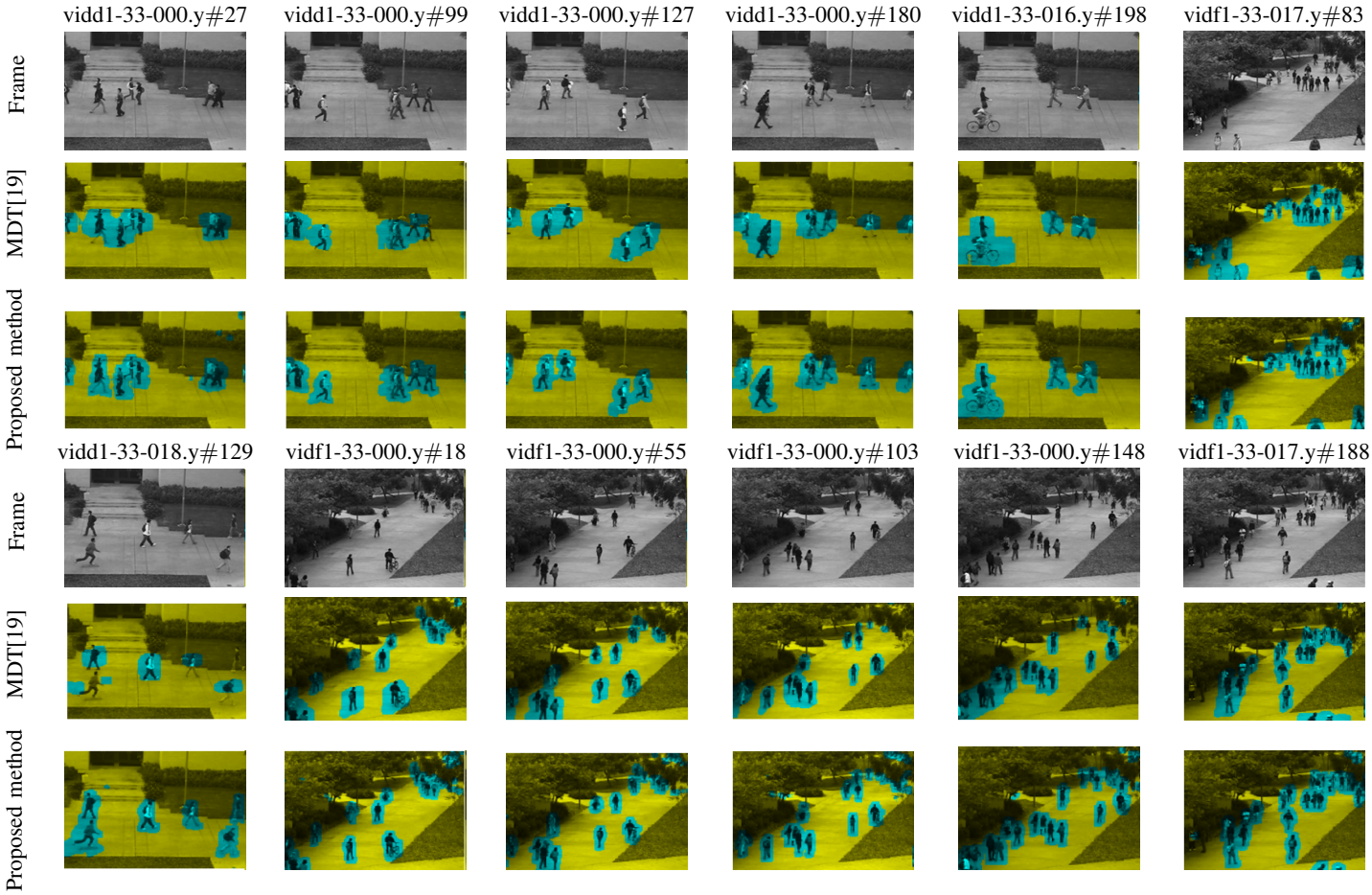

Fig. 11: Qualitative comparison of the motion detection results with MDT [19] for some frames of the sequences of the UCSD pedestrian dataset, (static segments colored yellow and motion segments colored green)

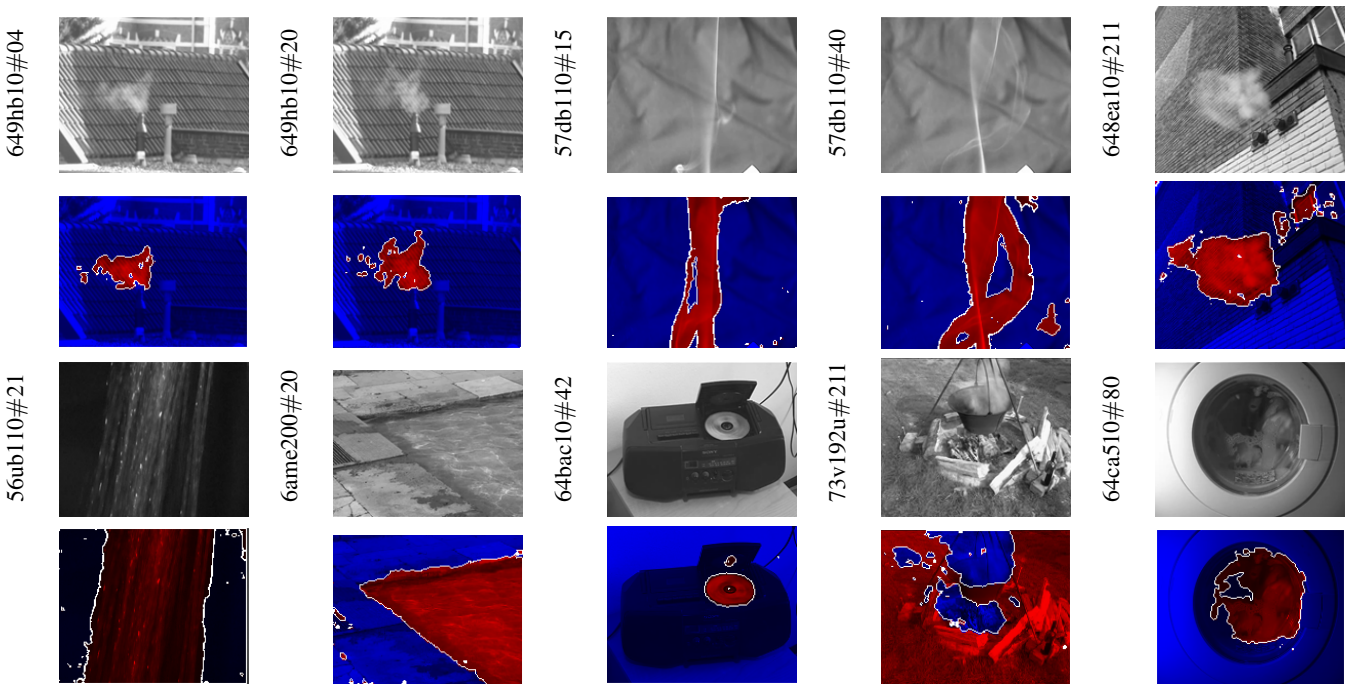

Fig. 12: Qualitative segmentation results of the proposed method for various frames of the video sequences in Dyntex dataset

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

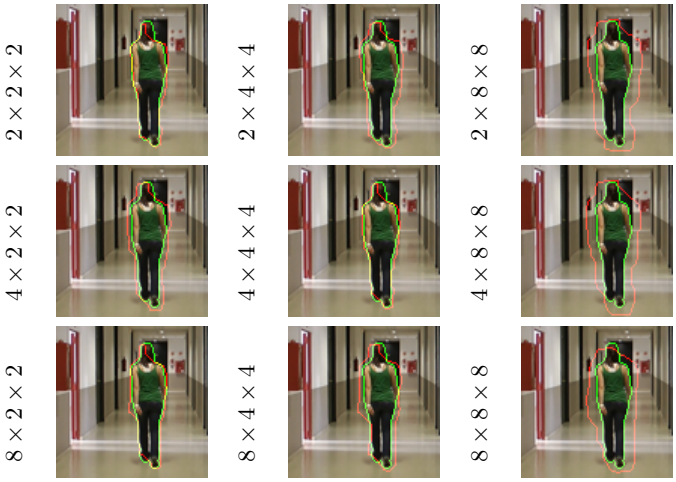

Fig. 13: Qualitative comparison of the different patch sizes for I_BS_01 sequence of LASIESTA dataset, green contour: ground truth and red contour: the output of the method.

are presented.

As motion is defined with respect to the temporal dimension, selecting a correct number of consecutive frames for motion detection is an important choice that we have to make. The experimental results in this paper was reported for a fixed patch size according to the best average of measures. The optimal number can however be different from one moving object in a video to another one. This value can be low for quick objects to reduce false positive predictions, and should be high for slow ones to reduce false negative predictions. Therefore, the quality of the proposed method is affected by the size of the patches. In our application, this issue has been alleviated by defining a user-tuned run-time constraint that demonstrates the patch size.

We have reported the results of the method using the K-means classifier. As this module is independent of the proposed feature extraction approach, it can be replaced with other classifiers. In the online available application (i.e. http://dspl.ce.sharif.edu/motiondetector.html), it is possible to choose a K-means classifier or a Gaussian Mixture Model classifier. A more robust method for this goal would consider the coherency in spatio-temporal dimensions by trajectory tracking through time (see [45]). Also, we did not consider camera jitter in this investigation. Generally, camera jitter can be introduced as a uniform additive motion noise affecting the trajectory of image features. Visentini et al. proposed a 2D wavelet transform based method for global camera motion detection [46]. Therefore, another point for further research is considering the camera motion.

VI. Conclusions

In this paper, we proposed a novel motion detection method using spatial frequency descriptors based on the three-dimensional wavelet transform and the three-dimensional wavelet leader. Due to the ability of frequency domain approaches in providing holistic motion pattern information, the proposed method can effectively deal with the difficulties raised by illumination changes, camouflage, and sudden motions. The proposed wavelet-based descriptors, can effectively be used for dynamic texture segmentation. Moreover, the proposed method had a good capability in detecting small moving objects. In order to evaluate the performance of the proposed method, various qualitative and quantitative comparisons were performed. Towards this goal, four different datasets i.e. LASIESTA (for finding the optimal patch size) and CD.net 2014, Dyntex, and UCSD pedestrian (for evaluation of the method) were used. Furthermore, various evaluation metrics were computed for each of these datasets. The results from these qualitative and quantitative comparisons demonstrated that the proposed approach outperforms existing methods, both in terms of motion detection and in the capability of segmenting the dynamic textures properly.

References

[1] Kunio Takaya. Detection of scene changes for video indexing by means of the mpeg motion vectors. In *ISPACS*, pages 447–450. IEEE, 2006.

[2] Lucia Maddalena and Alfredo Petrosino. A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection. *Neural Computing and Applications*, 19(2):179–186, 2010.

[3] Dong Liang, Shun'ichi Kaneko, Manabu Hashimoto, Kenji Iwatao, Xinyue Zhao, and Yutaka Satoh. Co-occurrence-based adaptive background model for robust object detection. In *AVSS*, pages 401–406. IEEE, 2013.

[4] Massimo De Gregorio and Maurizio Giordano. Change detection with weightless neural networks. In *CVPR*, pages 403–407, 2014.

[5] Xiqun Lu. A multiscale spatio-temporal background model for motion detection. In *ICIP*, pages 3268–3271. IEEE, 2014.

[6] Mohamed Sedky, Mansour Moniri, and Claude C Chibelushi. Spectral-360: A physics-based technique for change detection. In *CVPR*, pages 399–402, 2014.

[7] Rui Wang, Filiz Bunyak, Guna Seetharaman, and Kannappan Palaniappan. Static and moving object detection using flux tensor with split gaussian models. In *CVPR*, pages 414–418, 2014.

[8] Alina Miron and Atta Badii. Change detection based on graph cuts. In *IWSSIP*, pages 273–276. IEEE, 2015.

[9] Dong Liang, Manabu Hashimoto, Kenji Iwata, Xinyue Zhao, et al. Co-occurrence probability-based pixel pairs background model for robust object detection in dynamic scenes. *PR*, 48(4):1374–1390, 2015.

[10] Simone Bianco, Gianluigi Ciocca, and Raimondo Schettini. How far can you get by combining change detection algorithms? In *ICIAP*, pages 96–107. Springer, 2017.

[11] Pierre-Luc St-Charles, Guillaume-Alexre Bilodeau, and Robert Bergevin. A self-adjusting approach to change detection based on background word consensus. In *WACV*, pages 990–997. IEEE, 2015.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2885211, IEEE Transactions on Circuits and Systems for Video Technology

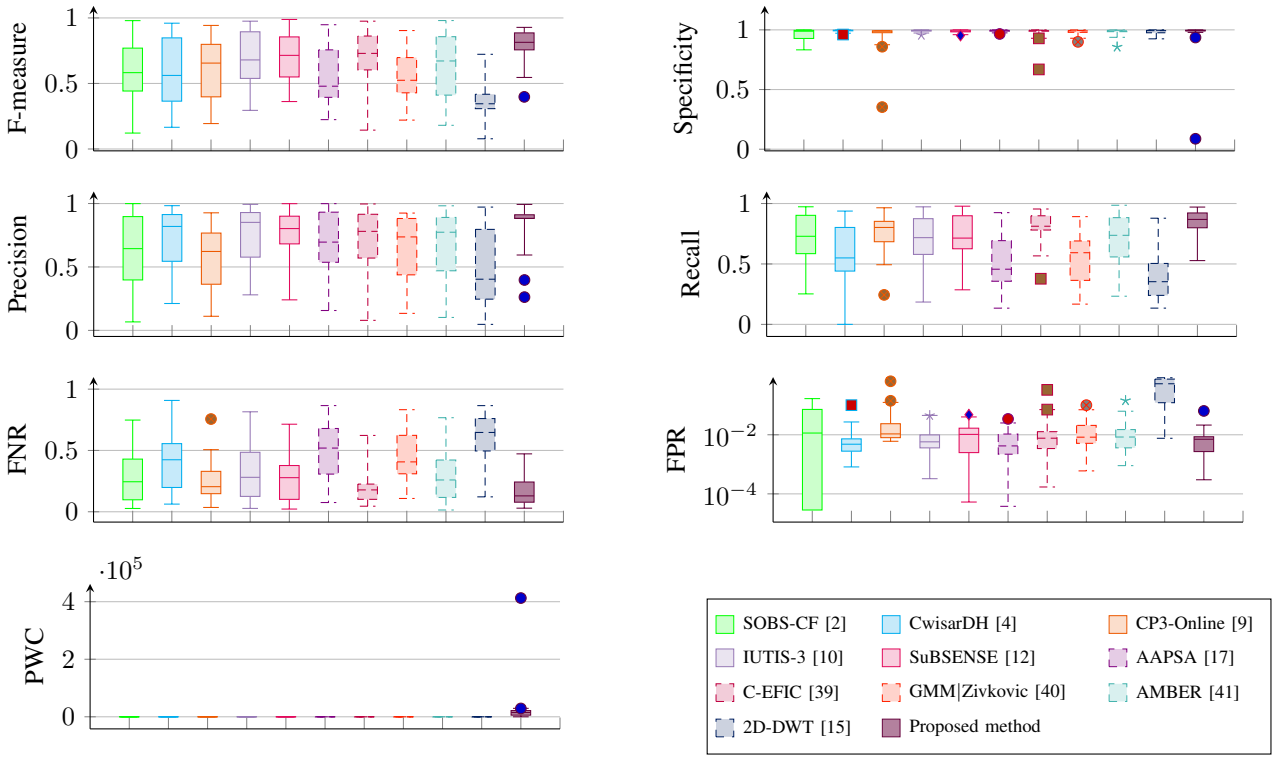

Fig. 14: Quantitative comparison of state-of-the-art with our proposed method using different measurements, according to the diagrams related to Recall, SP, Precision, and F-measure the 3D-DWT-MD outperforms the other methods sensibly.

[12] Pierre-Luc St-Charles, Guillaume-Alexre Bilodeau, and Robert Bergevin. Subsense: A universal change detection method with local adaptive sensitivity. *TIP*, 24(1):359–373, 2015.

[13] Martin Hofmann, Philipp Tiefenbacher, and Gerhard Rigoll. Background segmentation with feedback: The pixel-based adaptive segmenter. In *CVPRW*, pages 38–43. IEEE, 2012.

[14] Hui Ji, Xiong Yang, Haibin Ling, and Yong Xu. Wavelet domain multifractal analysis for static and dynamic texture classification. *TIP*, 22(1):286–299, 2013.

[15] J-C Huang and W-S Hsieh. Wavelet-based moving object segmentation. *Electronics Letters*, 39(19):1380–1382, 2003.

[16] B Ugur Töreyin, A Enis Cetin, Anil Aksay, and M Bilgay Akhan. Moving object detection in wavelet compressed video. *Signal Processing: Image Communication*, 20(3):255–264, 2005.

[17] Graciela Ramírez-Alonso and Mario I Chacón-Murguía. Auto-adaptive parallel som architecture with a modular analysis for dynamic object segmentation in videos. *Neurocomputing*, 175:990–1000, 2016.

[18] Pierre-Luc St-Charles, Guillaume-Alexandre Bilodeau, and Robert Bergevin. Universal background subtraction using word consensus models. *TIP*, 25(10):4768–4781, 2016.

[19] Antoni B Chan and Nuno Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *TPAMI*, 30(5):909–926, 2008.

[20] Renaud Péteri, Sándor Fazekas, and Mark J Huiskes. Dyntex: A comprehensive database of dynamic textures. *PRL*, 31(12):1627–1632, 2010.

[21] Feng Yang, Gui-Song Xia, Gang Liu, Liangpei Zhang, and Xin Huang. Dynamic texture recognition by aggregating spatial and temporal features via ensemble svms. *Neurocomputing*, 173:1310–1321, 2016.

[22] Kanoksak Wattanachote and Timothy K Shih. Automatic dynamic texture transformation based on a new motion coherence metric. *TCSVT*, 26(10):1805–1820, 2016.

[23] Wesley Nunes Gonçalves, Bruno Brandoli Machado, and Odemir Martinez Bruno. A complex network approach for dynamic texture recognition. *Neurocomputing*, 153:211–220, 2015.

[24] Stéphane Jaffard, Bruno Lashermes, and Patrice Abry. Wavelet leaders in multifractal analysis. In *Wavelet analysis and applications*, pages 201–246. Springer, 2006.

[25] Jayavardhana Gubbi, Slaven Marusic, and Marimuthu Palaniswami. Smoke detection in video using wavelets and support vector machines. *Fire Safety Journal*, 44(8):1110–1115, 2009.

[26] B Uğur Töreyin, Yiğithan Dedeoğlu, Uğur Güdükbay, and A Enis Cetin. Computer vision based method for real-time fire and flame detection. *PRL*, 27(1):49–58, 2006.

[27] Cédric Demonceaux and Djemâa Kachi-Akkouche. Motion detection using wavelet analysis and hierarchical markov models. *Lecture Notes in computer science*, 3667:64–75, 2006.

[28] Hansjörg Klock, Andreas Polzer, and Joachim M Buhmann. Region-based motion compensated 3d-wavelet transform coding of video. In *Image Processing, 1997. Proceedings., International Conference on*, volume 2, pages 776–779. IEEE, 1997.

[29] Zhi Li and Guizhong Liu. A novel scene change detection algorithm based on the 3d wavelet transform. In *ICIP 2008. 15th IEEE International Conference on*, pages 1536–1539. IEEE, 2008.

[30] Jizheng Xu, Zixiang Xiong, Shipeng Li, and Ya-Qin Zhang. Memory-constrained 3d wavelet transform for video coding without boundary effects. *TCSVT*, 12(9):812–818, 2002.

[31] Ivan W Selesnick, Jan E Odegard, and C Sidney Burrus. Nearly symmetric orthogonal wavelets with non-integer dc group delay. In *DSP*, pages 431–434. IEEE, 1996.

[32] A Farras Abdelnour and Ivan W Selesnick. Nearly symmetric orthogonal wavelet bases. In *ICASSP*, volume 6, 2001.

[33] Herwig Wendt, Patrice Abry, Stéphane Jaffard, Hui Ji, and Zuowei Shen. Wavelet leader multifractal analysis for texture classification. In *ICIP*, pages 3829–3832. IEEE, 2009.

[34] Xiaolin Chen, Xiaokang Yang, Shibao Zheng, Weiyao Lin, Rui Zhang, and Guangtao Zhai. New image quality assessment method using wavelet leader pyramids. *Optical Engineering*, 50(6):067011–067011, 2011.

[35] Nelly Pustelnik, Herwig Wendt, and Patrice Abry. Local regularity for texture segmentation: Combining wavelet leaders and proximal minimization. In *ICASSP*, pages pp–5348, 2013.

[36] Nelly Pustelnik, Herwig Wendt, Patrice Abry, and Nicolas Dobigeon. Local regularity, wavelet leaders and total variation based procedures for texture segmentation. Technical report, Tech. Rep, 2015.

[37] Hui Zhang, Jason E Fritts, and Sally A Goldman. A fast texture feature extraction method for region-based image segmentation. In *SPIE*, volume 5685, pages 957–968, 2005.

[38] Michel Marie Deza and Elena Deza. Encyclopedia of distances. In *Encyclopedia of Distances*, pages 1–583. Springer, 2009.

[39] Gianni Allebosch, David Van Hamme, Francis Deboeverie, Peter Veelaert, and Wilfried Philips. C-efic: Color and edge based foreground background segmentation with interior classification. In *VISIGRAPP*, pages 433–454. Springer, 2015.

[40] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *ICPR 2004*, volume 2, pages 28–31. IEEE, 2004.

[41] Bin Wang and Piotr Dudek. A fast self-tuning background subtraction algorithm. In *CVPR*, pages 395–398, 2014.

[42] Carlos Cuevas, Eva María Yáñez, and Narciso García. Labeled dataset for integral evaluation of moving object detection algorithms: Lasiesta. *Computer Vision and Image Understanding*, 152:103–117, 2016.

[43] Nil Goyette, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, and Prakash Ishwar. Changedetection. net: A new change detection benchmark dataset. In *CVPRW*, pages 1–8. IEEE, 2012.

[44] Haitao Guo and C Sidney Burrus. Fast approximate fourier transform via wavelets transform. In *Proc. SPIE Intl. Soc. Opt. Eng*, volume 2825, pages 250–259, 1996.

[45] Dirk Padfield, Jens Rittscher, and Badrinath Roysam. Spatio-temporal cell segmentation and tracking for automated screening. In *ISBI*, pages 376–379. IEEE, 2008.

[46] Marco Visentini-Scarzanella and Pier Luigi Dragotti. Video jitter analysis for automatic bootleg detection. In *MMSP, 2012*, pages 101–106. IEEE, 2012.

Sahar Yousefi is a Ph.D. candidate in artificial intelligence at the Sharif University of Technology, Iran. She has performed one-year research visiting at Leiden University Medical Center, The Netherlands. Her research interest is focused on image and video processing. She is also interested in machine learning, and probabilistic graphical models.

M.T. Manzuri Shalmani received his B.Sc. and M.Sc. degrees in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1984 and 1988, respectively. He received the PhD degree in electrical and computer engineering from Vienna University of Technology, Austria, in 1995. Currently, he is an associate professor in department of computer engineering, Sharif University of Technology, Tehran, Iran. His main research interests include digital signal processing, stochastic modeling, and Multi-resolution signal processing.

Jeremy Lin (SM 2005) received his M.S.E.E. from the University of Illinois, and a Ph.D. from the Drexel University. He is affiliated with PJM Interconnection, Audubon, PA. He was with GE Energy/Energy Consulting, Schenectady, NY, and Mid-America Interconnected Network Inc., Lombard, IL. His area of interest includes signal processing and optimization methods.

Marius Staring is an associate professor in biomedical machine learning at the Leiden University Medical Center, The Netherlands. His areas of research include machine learning for disease classification and staging, for segmentation, image registration and uncertainty estimation. Specific application areas encompass image analysis for neuro, lung and radiation therapy. He holds an MSc degree in Applied Mathematics from the University of Twente (2002), and a PhD degree from the UMC Utrecht on the topic of medical image registration (2008).