IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 17, 2018, accepted May 19, 2018, date of publication May 24, 2018, date of current version June 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2840329

Robustness Analysis of Pedestrian Detectors for Surveillance

YUMING FANG[1], (Senior Member, IEEE), GUANQUN DING[1], YUAN YUAN[2], WEISI LIN[2], (Fellow, IEEE), AND HAIWEN LIU[3], (Senior Member, IEEE)

[1]School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330013, China
[2]School of Computer Engineering, Nanyang Technological University, Singapore 639798
[3]School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Corresponding author: Yuming Fang (fa0001ng@e.ntu.edu.sg)

This work was supported in part by the National Natural Science Foundation of China under Grant 61571212, in part by the Fok Ying-Tong Education Foundation of China under Grant 161061, and in part by the Natural Science Foundation of Jiangxi under Grant 20071BBE50068 and Grant 20171BCB23048.

ABSTRACT To obtain effective pedestrian detection results in surveillance video, there have been many methods proposed to handle the problems from severe occlusion, pose variation, clutter background, and so on. Besides detection accuracy, a robust surveillance video system should be stable to video quality degradation by network transmission, environment variation, and so on. In this paper, we conduct the research on the robustness of pedestrian detection algorithms to video quality degradation. The main contribution of this paper includes the following three aspects. First, a large-scale distorted surveillance video data set (*DSurVD*) is constructed from high-quality video sequences and their corresponding distorted versions. Second, we design a method to evaluate detection stability and a robustness measure called *robustness quadrangle*, which can be adopted to the visualize detection accuracy of pedestrian detection algorithms on high-quality video sequences and stability with video quality degradation. Third, the robustness of seven existing pedestrian detection algorithms is evaluated by the built *DSurVD*. Experimental results show that the robustness can be further improved for existing pedestrian detection algorithms. In addition, we provide much in-depth discussion on how different distortion types influence the performance of pedestrian detection algorithms, which is important to design effective pedestrian detection algorithms for surveillance.

INDEX TERMS Object detection, video surveillance, image quality, video signal processing, image processing.

I. INTRODUCTION

Pedestrian detection plays an important role in auto-analysing of surveillance video. It is the prerequisite of various tasks of surveillance video processing including pedestrian tracking, crowd analysis, event recognition, anomaly detection, *etc*. During the last decade, significant progress has been achieved on existing published data sets including Caviar [1], INRIA [9], Caltech [12], PETS09 [17], TUD-Stadtmitte [3], *etc*. [45]. These data sets challenge the pedestrian detection algorithms by introducing different levels of occlusion, dynamic shape variation, different aspect ratios, *etc* [27]. By addressing these content-related challenges, various pedestrian detection algorithms [6], [8], [11], [28], [29], [33], [36], [37], [39], [40], [46] have been designed to obtain higher detection accuracy. It is reported [5] that the log-average miss-rate has decreased from around 70% [9] to around 35% [46] on Caltech data set [12].

However, in surveillance systems, the quality of surveillance video may change from time to time due to varies factors such as, bandwidth limitation, illumination variation, sensor variety of different cameras, *etc*. [18], [21]. When video quality decreases, targets may not be distinguishable any more in the distorted video, and this results in wrong detection. Thus, the effect of video quality variation on pedestrian detection should be investigated. There have been several studies focusing on assessing the quality of distorted image/video for face and event detection [19], [20]. It has been demonstrated that detectors always favor high quality image/video to obtain promising detection accuracy. On the other hand, a robust system requires detection algorithms which perform robustly and accurately in different quality conditions as well. There are some studies investigating into the benchmark of pedestrian detection [10], [12]. In these studies, the authors build a large-scale database to study the

FIGURE 1. Sample images in *DSurVD*, with four types of distortion introduced. Column (a) are the reference image frames with high quality; columns (b) and (c) are image frames with brightness variations; column (d) are image frames with additive white noise; column (e) are image frames with quality degradation after H.264 compression; column (f) are image frames with lower resolution.

statistics of the size, position and occlusion patterns of pedestrians in urban scenes. A new per-frame evaluation method is designed to measure the performance of different pedestrian detection algorithms. However, they do not consider the performance robustness of different pedestrian detection methods for quality-degradation video sequences. Currently, there is no systematic study focusing on the robustness of pedestrian detectors regarding to surveillance video quality variation, which motivates us to build a Distorted Surveillance Video Data Set (*DSurVD*) and study the robustness of pedestrian detectors to video quality degradation in surveillance systems. Our initial work has been reported in [43].

Generally, distortion in surveillance video may be caused by bandwidth limitation, noise in video acquisition, brightness variation due to camera variety, illumination change, etc. In this study, we consider the following four distortion types in the proposed *DSurVD*: compression distortion, resolution reduction, white noise and brightness changes. Regarding bandwidth limitation, distortion of video is mainly from compression distortion or resolution reduction. We also introduce different levels of white noise to the high-quality reference videos to obtain the noisy videos. Moreover, we adjust the brightness of the video to obtain the corresponding distorted versions with both high brightness and low brightness. Three common surveillance scenes are considered in the proposed *DSurVD*, including campus, town centre and car park. Fig. 1 illustrates some sample video frames in *DSurVD*.

Furthermore, we evaluate 7 existing pedestrian detectors which are published in studies [9], [11], [37], [39], [40] on the proposed *DSurVD*. To study the robustness of detectors, both the detection accuracy A_{ref} on high-quality reference videos and performance stability $\mathbf{S}$ on distorted videos are measured. With A_{ref} and $\mathbf{S}$, we define the robustness quadrangles (seen in Fig. 7) to visualize the robustness of different detectors. Based on the proposed robustness quadrangle, we know the advantages and disadvantages of detectors regarding to detection accuracy and stability with certain distortion type. Based on the in-depth analysis of the stability of existing pedestrian detectors with different distortion types, we facilitate some possibilities to improve the detection stability of pedestrian detectors regarding to video quality degradation.

The rest paper is organized as follows. We introduce the proposed *DSurVD* and the statistics of the distorted video sequences in Section II. Section III provides the definition of detection robustness; and a detection stability measurement is proposed (as Section III-B). In Section IV, the evaluation of pedestrian detectors on *DSurVD* is reported; the in-depth analysis is given in this section as well. Finally, we summarize the robustness study of pedestrian detection in surveillance video and discuss the possibilities in the future research in Section V.

II. DISTORTED SURVEILLANCE VIDEO DATA SET

In surveillance systems, due to bandwidth and storage limitation, video quality may vary after compression. Moreover, different environments, camera variety and unpredictable noise during video acquisition may influence the video quality as well. In this study, in order to study the performance of pedestrian detectors in surveillance video with quality variation, the first Distorted Surveillance Video Data Set (*DSurVD*) containing video sequences with distortion versions from different quality levels is constructed.

With H.264/AVC [38], one of the most widely used video coding standards for surveillance video, it is easy to adjust the *quantization parameter* (QP), *resolution* and frame rate to meet the limitation of the bandwidth. In general, if the QP and resolution are fixed, changing the frame rate does not significantly affect the quality of individual frames in a video sequence. In addition, since most pedestrian detectors process each video frame separately, without changing QP and resolution, frame rate does not affect detection accuracy. In the proposed *DSurVD*, video sequences with different

TABLE 1. Mean and coefficient of variation H_{vc} of pedestrian heights (in pixel); frame resolution of several existing pedestrian detection/tracking video data sets; the last row indicates whether the video sequences are captured by a fixed camera or not (See Section II-A for details).

	PETS09[17]	ParkingLot[34]	TownCentre[7]	TUD-Cam.[2]	Caviar[1]	Caltech[12]	ETH[14]
μ_h	83.8	171.6	203.6	207.5	64.1	51.4	254.2
H_{vc}	0.23	0.09	0.31	0.27	0.43	0.65	0.66
Res.	768×576	1920×1080	1920×1080	640×480	384×288	640×480	640×480
FixCam	$\surd$	$\surd$	$\surd$	$\surd$	$\times$	$\times$	$\times$

QPs and resolutions are created as distorted versions. Moreover, additive noise during video acquisition and brightness variation caused by illumination change or overexposing are two important distortion sources. Hence, two more distortion types of *white noise* and *brightness variation* are included with *DSurVD*.

A. REFERENCE VIDEO SEQUENCES

In *DSurVD*, the distorted video sequences are created based on five high quality surveillance video sequences including scenarios of campus (two sequences in PETS09 [17]), town centre (TownCentre sequence [7]) and car park (ParkingLot1 and ParkingLot2 sequences [34]). The ground truth are manually labeled bounding boxes of pedestrians.

These five video sequences are typical surveillance video data which have been widely used in recent pedestrian detection and tracking studies [7], [24], [35], [42], [44]. Furthermore, these sequences are with relatively high resolution and constant pedestrian size, and are captured with fixed cameras. Captured with fixed cameras guarantees relatively stable quality of all the frames in each video sequence. The reason why we prefer constant pedestrian size is as follows. As we reduce the resolution of reference videos, the lose of high frequency information of pedestrians caused by pedestrian size reduction is the main factor that affect the performance of detection algorithms. Thus, in order to study the relationship between the pedestrian size and detection accuracy, it is better to have a constant pedestrian size in the reference video. We use the variation coefficient of the pedestrian height (H_{vc}) to represent the variation of pedestrian size:

$$H_{vc} = \frac{\sigma_h}{\mu_h}, \tag{1}$$

where σ_h and μ_h denote the standard deviation and the mean of pedestrian height (in pixels) respectively, which are obtained from the ground truth of each data set. From Eq. 1, we can see that a smaller value of H_{vc} indicates more constant pedestrian size. Table 1 provides the H_{vc} and the resolution of some popular pedestrian detection/tracking video data sets. It shows that the five reference sequences used in *DSurVD* are with highly constant pedestrian size and relatively high resolution.

B. DISTORTED VIDEO SEQUENCES

For each reference sequence, we create 52 distortion versions (as explained next) based on the aforementioned four distortion types. Hence, including the reference sequences, there are $53 \times 5 = 265$ video sequences in total in the *DSurVD*. Below, we analyze the statistics of the *DSurVD* in detail.

1) QUANTIZATION PARAMETER

QP is one of the most important parameters in H.264 codec to encode video stream with different bit rates. The quality of the video is degraded by increasing the value of QP. In H.264, the Quantization Parameter (QP) determines the quantization step of the transformed coefficients with Discrete Cosine Transform. Larger QP refers to the bigger step and results in poorer video quality while lower QP refers to the smaller step and results in better quality. QP cannot directly refer to the bitrate since the bitrate is content biased. However, in general, each unit increase of QP lengthens the step size by 12% and reduces the bitrate by roughly 12% in H.264. Detailed information can be referred to the study [32]. In *DSurVD*, we encode each reference video sequence with 11 quality levels by varying QP from 10 to 65. These distorted sequences are named as SqsQP. The codec we used is ffmpeg [4].

The peak signal noise ratio (PSNR) of each distorted sequence is computed and shown in Fig. 2(b). Within the 11 distortion levels, PSNR drops from $\sim$ 50dB to $\sim$ 23dB. PSNR of PETS09L1 & 2 is a little lower than that of TownCentre and ParkingLot1 & 2, due to the distortion in the large grass regions which are with more complex texture in PETS09L1 & 2.

The compression ratio of each distorted sequence is computed and shown in Fig. 2(a). The compression ratio varies from 1 to around 10^3 between the reference sequences and the distorted sequences. The compression ratio of PETS09L1 is higher than that of PETS09L2 with the same QP even when they have the same background scene. The reason is the low density of pedestrians in PETS09L1. Based on the statistics of manually labelled ground truth, the average number of pedestrians per frame in PETS09L1 is 5.8 which is much lower than 23.6 in PETS09L2. Lower pedestrian density may result in less motion in the video, and thus less bite rate is required for inter coding between consecutive frames with H.264 codec.

2) RESOLUTION

Reducing the resolution is an alternative way to meet low bandwidth limitation in H.264 codec. In *DSurVD*, we code 11 video sequences with low resolution for each reference video sequence. These sequences are named as SqsRes.

To video sequences PETS09L1 & 2, the resolution is reduced from 768×576 (reference video) to 24×18.

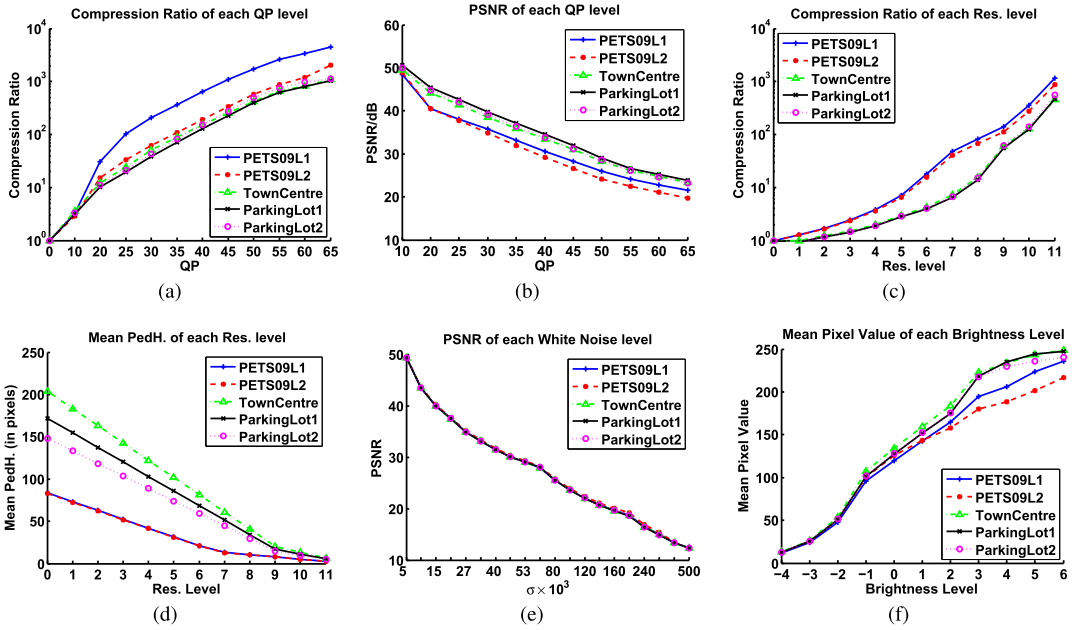


FIGURE 2. **(a) and (b) are the compression ratio and PSNR statistics of the distorted videos by adjusting the QP; (c) and (d) are the compression ratio and mean pedestrian height of the distorted videos by adjust the resolution, higher resolution level indicates smaller resolution of the video and 0 refers to the reference videos; (e) is the PSNR statistics of the distorted videos by adding zero mean Gaussian noise, where x-axis is the σ of the Gaussian kernel; (f) is the mean pixel value of distorted video with brightness variation, "0" indicates the reference video.(See Section II-B for details).**

For video sequences TownCentre and ParkingLot1 & 2, the resolution is reduced from 1920 × 1080 (reference video) to 64 × 36. The compression ratio of each distorted sequence is computed and shown in Fig. 2(b). "0" in the Res. Level axis indicates the reference video sequences and "11" indicates the video sequences with the lowest resolution (24 × 18 for PETS09L1 & 2, and 64 × 36 for TownCentre and ParkingLot1 & 2).

Pedestrian size change is the direct effect by reducing the resolution of video sequences. As mentioned in [12], pedestrian height cannot be neglected for pedestrian detection accuracy. We plot the relation between the average pedestrian height and the resolution level of each sequence in Fig. 2(d). The average pedestrian height (in pixels) varies from around 200 to about 10 in the proposed *DSurVD*. From level "0" to level "7", the down sampling step of each video is kept the same, and it can be seen from Fig. 2(d) that the slope of the each line keeps the same before level "7". In order to study more detail of pedestrian detectors with low resolution (or small pedestrian size), we decrease the down sampling step after level "7" to get more low resolution videos, and we can see that the slope of each line become smaller after level "7".

3) WHITE NOISE

Apart from the aforementioned distortion types introduced by compression, white noise is another common type of distortion during image/video acquisition [30]. We use the zero-mean Gaussian noise to model the additive white noise. In total, 20 levels of Gaussian noise are added to the reference video sequences where the Gaussian Kernel σ varies from 0.005 to 0.5, and the PSNR varies from $\sim$ 50dB to $\sim$ 25dB, respectively. These sequences are named as SqsWN. The PSNR of each noisy sequence is computed and shown in Fig. 2(e). The PSNR is highly correlated to σ and there is almost no difference of PSNR between different reference video sequences.

4) BRIGHTNESS VARIATION

The brightness variation of video frames in surveillance video sequences can be caused by both illumination change and different exposure sensitivity of the camera. We model 10 levels of brightness for video sequences in *DSurVD*. These sequences are named as SqsBV. In total, the distortion versions with 4 low and 6 high brightness levels are created. The mean pixel value of each brightness level are shown in Fig. 2(f).

III. ROBUSTNESS ANALYSIS: ACCURACY AND STABILITY

The IEEE Standard Glossary of Software Engineering Terminology [31] gives the definition of robustness as follows:

Robustness: *The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions.*

Based on this definition, the robustness of the pedestrian detector can be measured from the following two aspects. On one hand, what is the accurate rate of the pedestrian detector in surveillance videos? On the other hand, what is the performance stability of the pedestrian detector with video quality degradation? In this section, we denote these

two aspects as accuracy and stability, respectively, and propose an approach to measure them.

A. ACCURACY MEASUREMENT

Given a detection bounding box bb_{dt} and a ground truth bounding box bb_{gt}, we employ the matching criterion used in the previous pedestrian detection benchmark study [12]. With the overlap between bb_{dt} and bb_{gt} exceeding 50 percent, we consider them as a correct match,

$$\frac{area(bb_{dt} \bigcap bb_{gt})}{area(bb_{dt} \bigcup bb_{gt})} > 0.5. \tag{2}$$

Each bb_{gt} can match to at most one bb_{dt}. If a detection bounding box matches multiple ground truth bounding boxes, the match with the highest overlap is used. Unmatched bb_{dt} and bb_{gt} are considered as false positive samples and false negative samples, respectively.

We plot miss rate against false positives per image (FPPI) to visualize the performance of pedestrian detector (e.g., Fig. 4). Similar to [12], the *log-average miss rate* ($\mathcal{MR}$) is computed by averaging miss rate at nine FPPI rates evenly spaced in log-space in the range 10^{-2} to 10^{0}, to quantify the performance of the detector. We define the *Accuracy* by subtracting $\mathcal{MR}$ by 1 to make sure it is positive related to performance:

$$A = 1 - \mathcal{MR}. \tag{3}$$

The value of A ranges from [0, 1] and larger value indicates better performance of pedestrian detector.

B. STABILITY MEASUREMENT

Accuracy is an important measurement index for pedestrian detectors. In traditional pedestrian detection data sets [15], different *Accuracy* metrics have been proposed to measure the performance of pedestrian detectors. However, *Stability* is another unneglectable measurement index for pedestrian detectors. In the case where two pedestrian detectors have similar *Accuracy* on good quality-video sequences, the *Stability* measurement provides another important dimension to evaluate the performance. Here, we propose a *Stability* (**S**) measurement method by analyzing the performance of pedestrian detectors in surveillance videos with quality degradation. To the best of our knowledge, this is the first study to provide the dedicated analysis of stability of pedestrian detectors for visual surveillance with video quality variation.

With the four aforementioned common distortion types, we define *Stability* as a four dimensional vector,

$$\mathbf{S} = [\mathbf{S}.\text{qp}, \mathbf{S}.\text{res}, \mathbf{S}.\text{wn}, \mathbf{S}.\text{bv}], \tag{4}$$

where **S**.qp, **S**.res, **S**.wn and **S**.bv denote the *Stability* of pedestrian detectors with QP variation, resolution variation, additive white noise and brightness variation, respectively.

To quantify the *Stability*, two criteria are incorporated in the study as follows:

1) RATE OF ACCURACY DEGRADATION

The *Accuracy* degradation rate of a robust detector should be slow when input video quality decreases.

2) MONOTONICITY

A robust detector would show a monotonically degradation in *Accuracy* when input video quality decreases.

It is easy to understand the slow accuracy degradation rate criterion in degradation study. The motivation of the monotonicity criterion is that, with quality degradation, detectors whose detection accuracy oscillates are much less predictable than detectors with monotonically accuracy degradation. In other words, when increasing the video quality gradually, we prefer monotonically increasing of the detection accuracy rather than oscillating of the detection accuracy.

Given a reference video sequence and a particular distortion type x (e.g., x can be qp, res, wn and bv), we first compute the detection *Accuracy* values with the reference video sequence A_{ref} and all the distorted video sequences $\{A_i : i = 1, \ldots, N_{\text{x}}\}$, where N_{x} is the number of distorted sequences with distortion type x.

For the i^{th} distorted sequence, we formulate the penalty of accuracy degradation PD_i:

$$PD_i = \min\{1, (\frac{A_i - A_{\text{ref}}}{A_{\text{ref}}})^2\}, \tag{5}$$

where A_i, and A_{ref} are the detection accuracy on the i^{th} distorted video sequences and the reference video sequence, respectively. It can be seen that in Eq. 5, PD_i is positive correlated to the difference between A_i and A_{ref}. In other words, less penalty will be assigned if A_i is more closer to A_{ref}. PD_i ranges from 0 to 1. If A_i is much greater than A_{ref} (e.g., $A_i > 2A_{\text{ref}}$) which rarely happens in the robustness test, we limit the penalty to be 1. Actually, the penalty of accuracy degradation PD_i describe the invariance property of the detection accuracy.

Furthermore, the non-monotonicity penalty of the i^{th} distorted sequence PM_i is formulated as:

$$PM_i = \begin{cases} 0 & A_i \leq A_{i-} \\ \min\{1, (\frac{A_i - A_{i-}}{A_{i-}})^2\} & A_i > A_{i-}, \end{cases} \tag{6}$$

where A_i, A_{i-}, and A_{ref} are the detection accuracy on the i^{th}, $i-^{th}$ distorted video sequences and the reference sequence, respectively.

Here, we give the definition of the $i - th$ distorted video sequence. With distortion type of qp, res or wn, we simply rank the distorted sequences with quality descending. The $(i-)^{th}$ distorted sequence in Eq. 6 is just next to the i^{th} one and with better quality.[1] With distortion type of bv, the distorted sequences are divided into high brightness sequences and low brightness sequences comparing with the reference video. By ranking these two groups separately with quality

[1] If $i = 1$, $(i-)^{th}$ is the reference video sequence.

descending, the $(i-)^{th}$ distorted sequence is next to the i^{th} one in the same group with better quality.

To meet the monotonicity criterion, A_i is supposed to be not larger than A_{i-} since the quality of the i^{th} distorted video sequence is worse than that of the $(i-)^{th}$ distorted video sequence. Thus, if $A_i > A_{i-}$, penalty will be assigned as shown in Eq. 6. With the same concern in Eq. 5, if A_i is much greater than A_{i-} (e.g., $A_i > 2A_{i-}$), we limit the penalty to be 1.

Based on these two penalty functions, we compute the *Stability* with given distortion type x (e.g., x can be qp, res, wn and bv) as:

$$\mathbf{S}.\mathrm{x} = 1 - \sqrt{\frac{1}{N_\mathrm{x}} \sum_{i=1}^{N_\mathrm{x}} \omega PD_i + (1-\omega) PM_i}, \tag{7}$$

where ω is the weighting parameter between PD_i and PM_i, which is used to adjust the importance of these two factors, and it ranges from 0 to 1.

Fig. 3 shows the flowchart of computing *Stability*. The quantified *Stability* ranges from 0 to 1, and a higher value shows more stable performance of a pedestrian detector. The stability value of an ideal stable detector should be 1 with $A_i \triangleq A_{\mathrm{ref}}$.

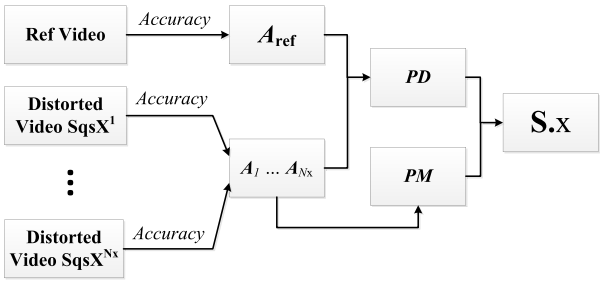


FIGURE 3. ***Stability*** **evaluation based on the variability of** ***Accuracy*** **from the reference video to the most distorted video. SqsX^1 to SqsX^N indicate the distorted video sequences with distortion type x.**

IV. EVALUATION RESULT

In this section, we show the evaluation result of several existing pedestrian detectors on *DSurVD*.[2] The comparison and robustness between the tested detectors are given, along with the discussion about the evaluation result.

A. EVALUATED DETECTORS

We evaluate the performance of seven representative existing pedestrian detectors (Table 2) on *DSurVD*. The source codes of these detectors are obtained from their corresponding public websites. To evaluate the detectors, we use the pre-trained pedestrian model and the default parameter values provided by the authors. We believe that this is fair since the authors have the best knowledge in tuning the parameters. Here, we give some brief introduction of these evaluated detectors, while a thorough survey of pedestrian detectors can be referred to [5] and [13].

As a type of gradient-based features, histogram of gradient (HOG) [9] shows substantial gains over conventional intensity-based features. The evaluated HOG pedestrian detector is based on the sliding window paradigm, while a soft linear SVM classifier is used to classify the positive pedestrian windows and negative windows. Compared with HOG pedestrian detector, HogLbp [37] pedestrian detector uses the combination of HOG and local binary pattern (Lbp) as the feature descriptor. By representing the edge/local shape information, Lbp feature is a complement to HOG feature when the background is cluttered with noisy edges. Moreover, the integration of global and part-based detectors in HogLbp pedestrian detector improves the detection accuracy when the targets are partial occluded. In the study of LatSVM [16], a deformable part-based pedestrian detector is designed. The unknown part positions are modeled as latent variables in an SVM framework, which allows reasonable deformation of the target. Moreover, a PCA-HOG (Principal Component Analysis-HOG) is proposed to reduce the feature dimensionality with no noticeable loss of information. The difference between LatSVM and LatSVM-IN in Table 2 is the training data, LatSVM is trained with the PASCAL training data while LatSVM-IN is trained with INRIA training data. In C4 [40] pedestrian detector, a cascade classifier with two nodes of linear SVM and Histogram Intersection Kernel (HIK) SVM [23] is used. And in order to explicitly encode the human contour information, the CENTRIST feature [41] is used in C4. As reported in [5], Aggregate Channel Feature (ACF) detection framework achieves state-of-art performance in pedestrian detection. HOG and LUV color features are used, and boosted trees are trained in ACF. The difference between ACF and ACF-Cal in Table 2 is the training data, ACF is trained with INRIA training data while ACF-Cal is trained with Caltech training data.

TABLE 2. **List of tested pedestrian detectors. (ACF, ACF-Cal) and (LatSVM, LatSVM-IN) are the same algorithms but with different training data. The pedestrian model height of each detector is measured in pixels.**

Detectors	Feature Type	Classifier	Training Data	Model Height
HOG[9]	HOG	linear SVM	INRIA	96
HogLbp[37]	HOG+Lbp	linear SVM	INRIA	96
LatSVM[16]	PCA-HOG	latent SVM	PASCAL	80
LatSVM-IN[16]	PCA-HOG	latent SVM	INRIA	96
C4[40]	CENTRIST	linear + HIK SVM	INRIA	108
ACF[11]	HOG+LUV	AdaBoost	INRIA	100
ACF-Cal[11]	HOG+LUV	AdaBoost	Caltech	50

B. ACCURACY ON REFERENCE VIDEO SEQUENCES

The *Accuracy* of detectors on the reference video sequences reflects the performance of detectors on good-quality videos.

[2]Only results on sequences PETS09L1, TownCentre and ParkingLot1 are shown in this paper due to the page limitation. The full results can be achieved on: https://sites.google.com/site/sorsyuanyuan/home/rdetection

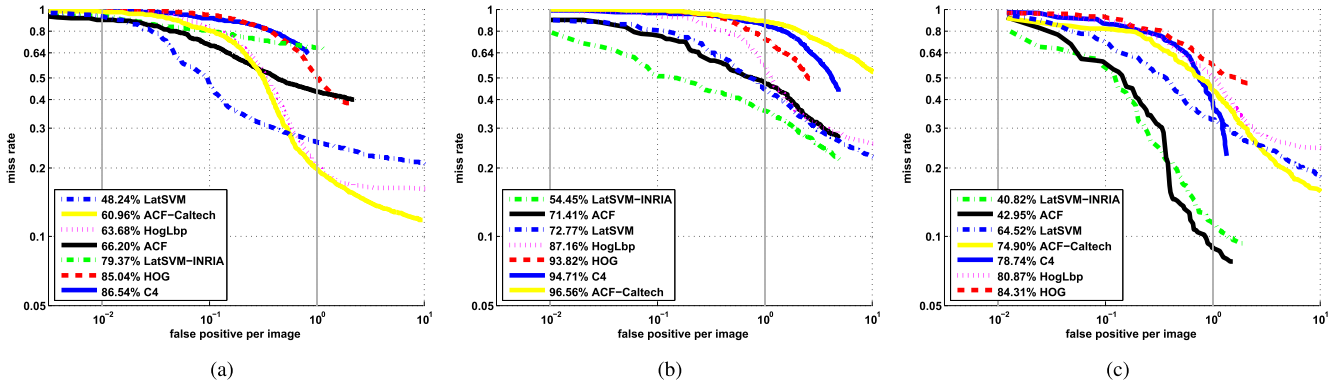


FIGURE 4. Miss rate vs. FPPI on reference video sequences (See Section IV-B for details). (a) PETS09L1. (b) TownCentre. (c) ParkingLot1.

The miss rate vs. FPPI of detectors on reference video sequences are plotted in Fig. 4. Legend entries show the log-average miss rate of each detector from best to worst. It can be seen that LatSVM (or LatSVM-IN) and ACF (or ACF-Cal) consistently perform better compared with other algorithms. On PETS09L1, LatSVM and ACF-Cal show higher detection accuracy than LatSVM-IN and ACF respectively, while on TownCentre and ParkingLot1, LatSVM-IN and ACF perform better. From Table 2, it can be seen that the model heights (in pixels) of LatSVM (80) and ACF-Cal (50) are smaller than LatSVM-IN (96) and ACF (100) respectively. And from Table 1, it can be seen that the pedestrian height of PETS09L1 is smaller than that of TownCentre and Parkinglot1 due to smaller resolution. We find that to the same algorithm, training with small model height gives better performance on sequences with small pedestrian height, and vice versus.

C. QUADRANGLE: ROBUSTNESS REPRESENTATION

With the definition in Sec. III, the robustness of a detector can be described by a combination of *Accuracy* (A_{ref}) on good-quality video and *Stability* (**S**) with four types of distortions. The detection accuracy of detectors on distorted video sequences are computed and plotted in Fig. 6. Besides, the performance of *Stability* with different weighting parameters ω are plotted in Fig. 5. From Fig. 5 and Fig. 6, it can be seen that most detectors follow the monotonicity criterion when the video quality drops. Additionally, with different weighting parameters ω, the ranking order of *Stability S* from different pedestrian detectors are almost kept stable, which demonstrates that the adjustment of parameter ω has little effect on the ranking order of *Stability S* by different pedestrian detection algorithms. In other words, with larger parameter ω, the performance of the pedestrian detection methods decreases when the video quality drops. However, there might be some pedestrian detector with bad monotonicity in the literature. Thus, we hold the second factor with small value to provide good extensibility for the proposed metric. This is the reason why we assign more weighting to the penalty of accuracy degradation *PD* by setting $w = 0.8$ in Eq. (7) when computing the detection performance *Stability*. In the experiment, we found that the detection accuracy would not decrease greatly with video quality dropping on a robust pedestrian detector, which demonstrates that the initial hypothesis of monotonicity is convincing. In the future, we will further investigate into this weighting parameter to design better metric.

When comparing the robustness of two detectors, we first consider the value of A_{ref}. If there exists a large A_{ref} difference between the compared detectors, the detector with the larger A_{ref} is more preferable and we should take less consideration on the *Stability*. On the other hand, if the compared detectors are with a similar value of A_{ref} (this is likely the case for the relevant state-of-the-art detectors), then **S** becomes an important criterion to robustness.

Based on this analysis, we propose the robustness quadrangle (as shown in Fig. 7) to visualize A_{ref} and **S**. For each quadrangle, the heights of four angles represent the *Stability* to the four types of distortion respectively (**S**.qp, **S**.res, **S**.wn, **S**.bv). The center point of each quadrangle indicates $A_{\text{ref}} * \lambda$ where λ is a scaling factor of A_{ref} in robustness and decides the range of x-axis of the robustness quadrangles figure.

By given a non-zero value to λ (e.g. $\lambda = 5$, $\lambda = 2$), the robustness quadrangles of evaluated detectors are shown in the left and middle column of Fig. 7. If two detectors are with large A_{ref} difference, we can intuitively read the A_{ref} difference based on the distance between two robustness quadrangles. If the A_{ref} values of two detectors are similar, the center points of two quadrangles are close and we can straightforward compare their **S** values by the corners of two overlapped quadrangles. The red square on the most right hand side with dashed boundary represents the ideal detector whose $A_{\text{ref}} = 1$ and $\mathbf{S} = [1, 1, 1, 1]$.

If we want to emphasis A_{ref} more in the robustness quadrangle figure, we can set larger λ (e.g. $\lambda = 5$, the left column of Fig. 7). Thus the differences of A_{ref} between detectors will be amplified on the x-axis. If we want to emphasis **S** more in the robustness quadrangle figure, we can set smaller λ with the same reason (e.g. $\lambda = 2$, the middle column of Fig. 7). $\lambda = 0$ is an extreme case that we only compare **S** of detectors

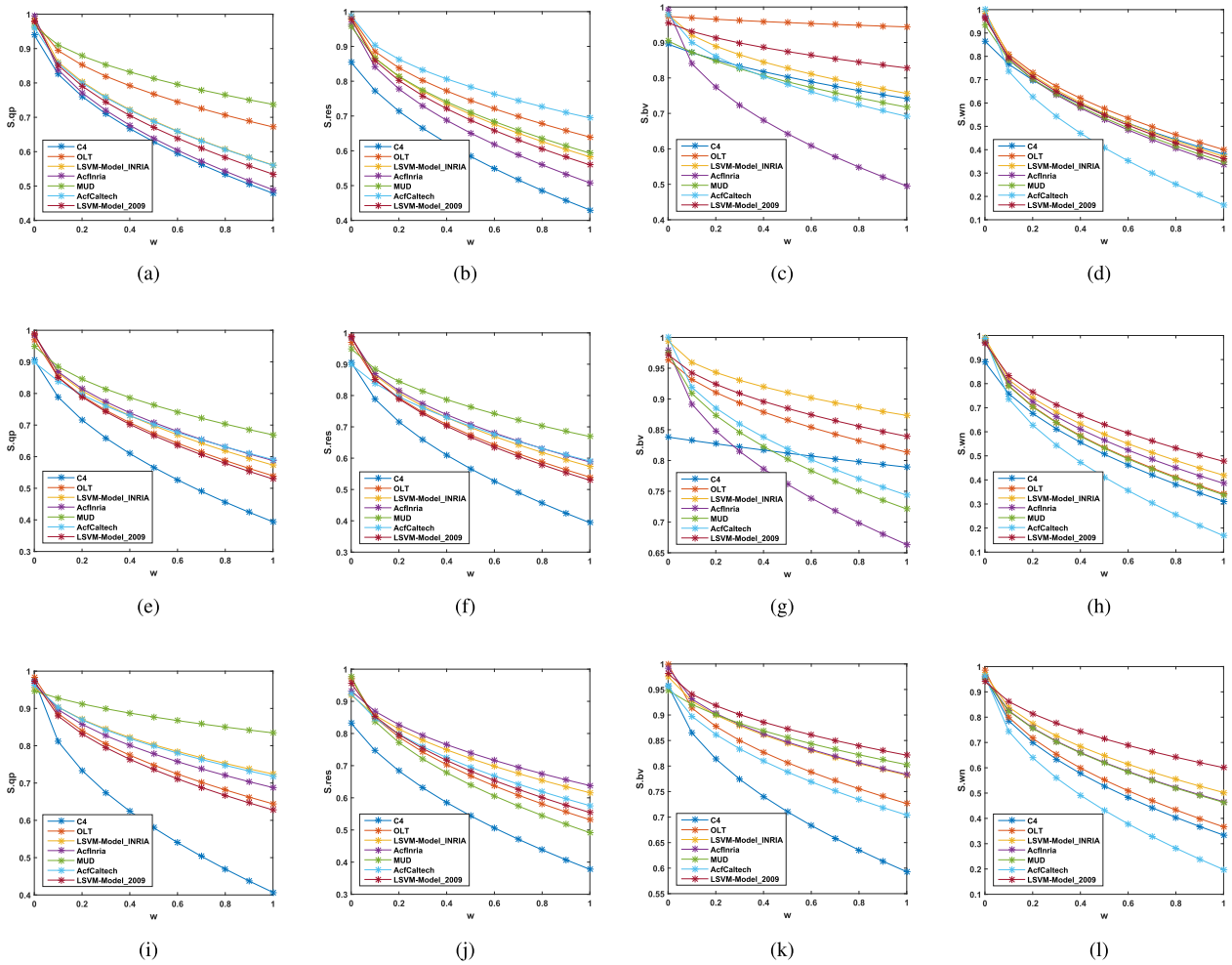


FIGURE 5. The performance of *Stability S* with different weighting parameters ω on different datasets. X-axis refers to the different weighting parameters ω in the range from 0 to 1, and y-axis represents the performance of *Stability S* with four attributions (S.qp: Quantization Parameters, S.res: Resolution, S.bv: Brightness Variation, S.wn: White Noise). (a) S.qp with ω on TownCentre. (b) S.res with ω on TownCentre. (c) S.bv with ω on TownCentre. (d) S.wn with ω on TownCentre. (e) S.qp with ω on ParkingLot1. (f) S.res with ω on ParkingLot1. (g) S.bv with ω on ParkingLot1. (h) S.wn with ω on ParkingLot1. (i) S.qp with ω on ParkingLot2. (j) S.res with ω on ParkingLot2. (k) S.bv with ω on ParkingLot2. (l) S.wn with ω on ParkingLot2.

while all the center points of quadrangles converge to $[0, 0]$ and we cannot see any difference between A_{ref}. The right column of Fig 7 shows the the robustness quadrangles with $\lambda = 0$.

D. STABILITY WITH QP VARIATION

The first row of Fig. 6 shows the *Accuracy* vs. PSNR curves of evaluated detectors with three different scenes (Campus, Town Centre and Car Parking). It can be seen that, in general, the *Accuracy*-PSNR curves of most detectors are monotonic. And the degradation of *Accuracy* of detectors is not obvious before some critical PSNR points (e.g., before 35dB). One noticeable fact is that, to each detector, the *Accuracy* fluctuates around A_{ref} before it dramatically decreases. This is because when the pixel values slightly changes due to compression, it would affect the decision of the algorithms on detections near to the threshold.

From the *Stability* values of left-upper corners in Fig. 7 and the ranking in Fig. 6(a), 6(b) and 6(c), HogLbp is most stable with QP variation among the evaluated detectors by ranking 1st on both TownCentre and ParkingLot1 and 2nd on PETS09L1. Comparing to HOG, the main modification of HogLbp is the feature type. Hence, Lbp feature [26] can be an important complement to HOG feature in pedestrian detection on heavily compressed surveillance videos with large QP. Moreover, with QP variation, ACF-Cal shows higher stability than ACF, and LatSVM-IN shows higher stability than LatSVM. This indicates that the training data is another important factor to detection stability. With the same algorithm of ACF, the detector trained with Caltech [12] data set performs more stable than the detector trained with INRIA [9] data set; with the same algorithm of LatSVM, the detector trained with INRIA data set performs more stable than the detector trained with PASCAL [15]. More exploration is needed toward

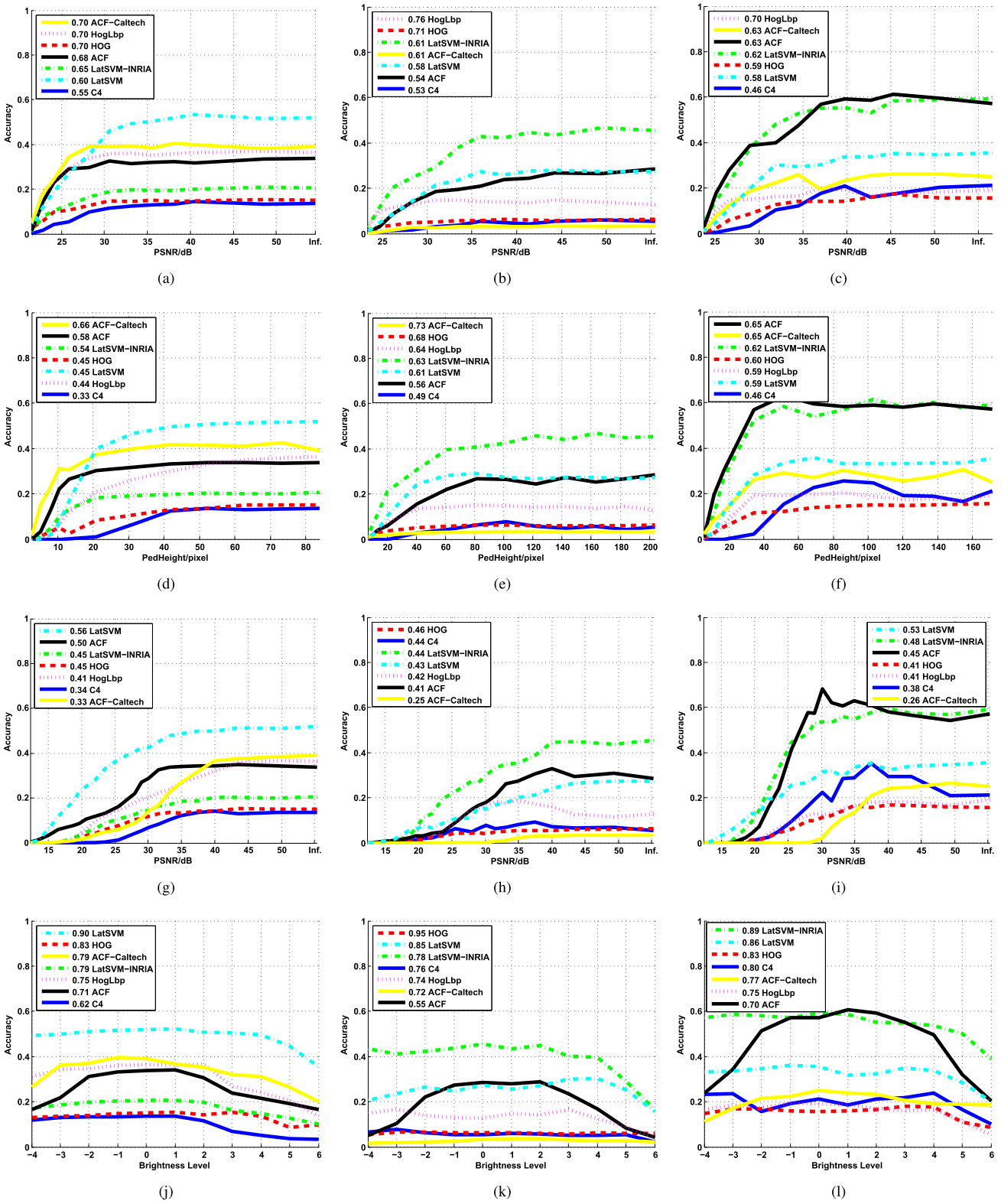


FIGURE 6. Detection accuracy variation of evaluated detectors with different types of distortion. In the first row and third row, PSNR=Inf. refers to the reference video sequences with the best quality. In the second row, the x-axis refers to the average pedestrian heights of the tested sequences. In the fourth row, the x-axis indicates the brightness level. "0" denotes the reference video sequences while negative values denote low brightness sequences and positive values denote high brightness sequences. (a) S.qp on PETS09L1. (b) S.qp on TownCentre. (c) S.qp on ParkingLot1. (d) S.res on PETS09L1. (e) S.res on TownCentre. (f) S.res on ParkingLot1. (g) S.wn on PETS09L1. (h) S.wn on TownCentre. (i) S.wn on ParkingLot1. (j) S.bv on PETS09L1. (k) S.bv on TownCentre. (l) S.bv on ParkingLot1.

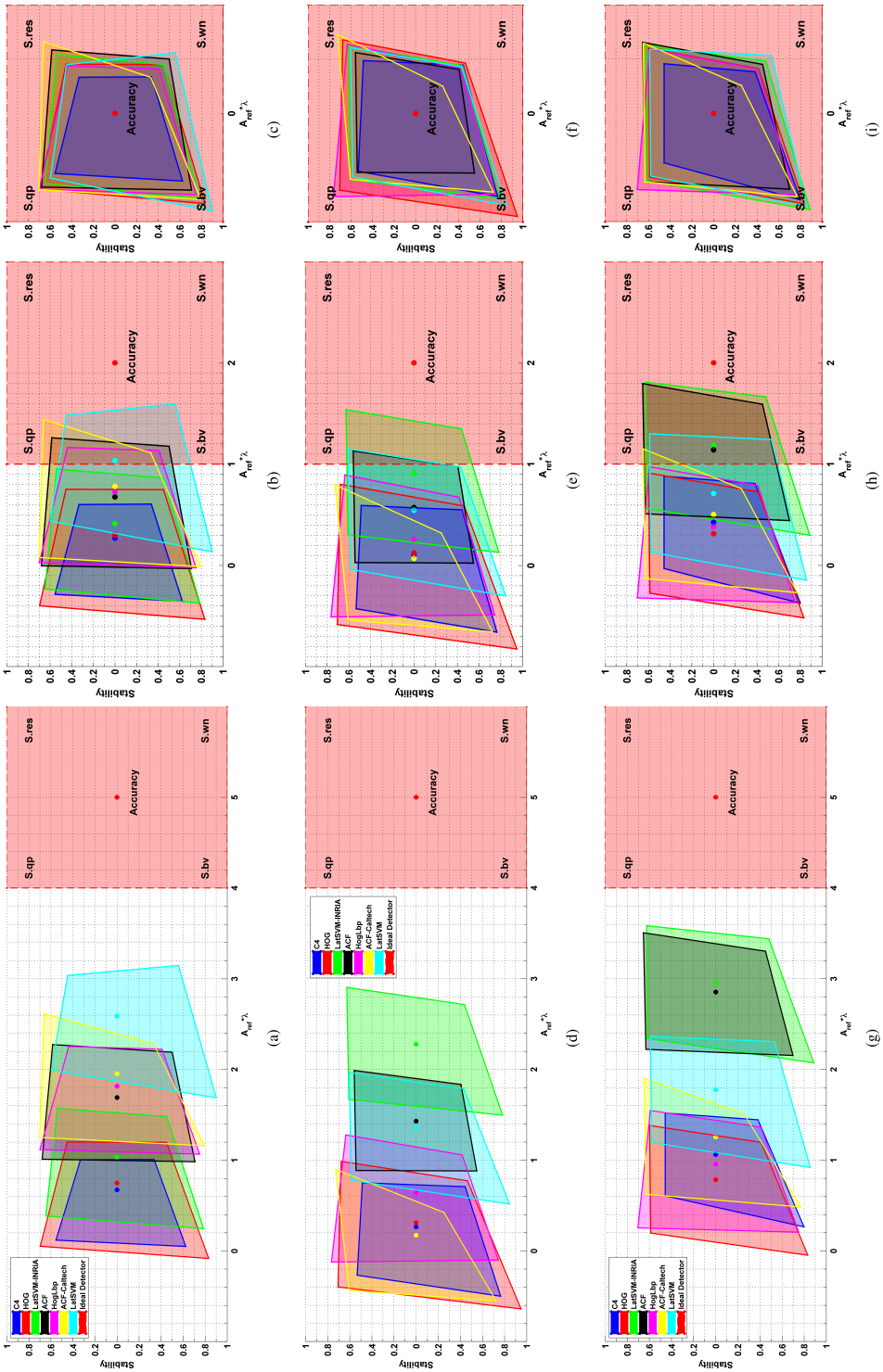


FIGURE 7. Robustness Quadrangles on tested sequences. The center location of each quadrangle read from x-axis denotes $A_{\text{ref}} * \lambda$ of each detector. The heights of four corners read from y-axis denote stability values with four types of distortions, respectively (QP variation, Resolution variation, White noise and Brightness variation from left-upper corner to left-bottom corner clockwise). For the ideal detector which is plotted in red color with dashed line, the center of corresponded quadrangle is $1 * \lambda$ and the heights of all the four corners are 1. When comparing the robustness of two detectors, the center locations of quadrangles are first compared and quadrangle on the right side is preferred. When the center locations of two quadrangles are close, the differences of stability values with different distortion types can be easily read from the corner heights. λ is a scaling factor of A_{ref} in robustness measurement, by setting $\lambda = 0$ (shown in the right column), only the differences of stability values can be read and the differences between A_{ref} are obscured. (See Section IV-C for details). (a) PETS09L1, $\lambda = 5$. (b) PETS09L1, $\lambda = 2$. (c) PETS09L1, $\lambda = 0$. (d) TownCentre, $\lambda = 5$. (e) TownCentre, $\lambda = 2$. (f) TownCentre, $\lambda = 0$. (g) ParkingLot1, $\lambda = 5$. (h) ParkingLot1, $\lambda = 2$. (i) ParkingLot1, $\lambda = 0$.

generalization of learning-based approaches and tackle overfitting.

E. STABILITY WITH RESOLUTION VARIATION

To resolution degradation of video sequences, pedestrian height change has the most direct impact which affects the detection accuracy. We plot the *Accuracy* vs. mean Pedestrian Height (in pixels) curves of evaluated detectors in the second row of Fig 6. It can be seen in most of the cases, videos with larger pedestrian height are with higher *Accuracy* values. Also, before the pedestrian height decreases to some critical points (e.g., 40 to PETS09L1, and 60 to TownCentre and ParkingLot1), the *Accuracy* values of detectors are relatively stable with pedestrian height change.

From the *Stability* values of right-upper corners in Fig. 7 and the ranking in Fig. 6(d), 6(e) and 6(f), we can see that ACF-Cal performs more constantly even on videos with low pedestrian height. Note that, the ACF-Cal is with the lowest model height (50) for training among the evaluated detectors. It reveals that the lower model height for training can result in stable performance to videos with low pedestrian height (or low resolution). However, a detector with too low model height may cause low *Accuracy* in detection as well (see the performance of ACF-Cal on TownCentre). We should be careful in defining the model height in the later detector designing to achieve both high *Accuracy* and *Stability*.

F. STABILITY WITH ADDITIVE WHITE NOISE

The third row of Fig. 6 shows the *Accuracy* vs. PSNR curves of evaluated detectors with different levels of additive white noise. It can be seen that, for most detectors, there is no obvious degradation of the *Accuracy* before some critical PSNR (e.g., before 40dB for the cases being studied) points. For noisy sequences of PETS09L1 and ParkingLot1, LatSVM and LatSVM-IN shows preferable stability comparing with other detectors based on the *Stability* values of right-bottom corners in Fig. 7 and ranking in Fig. 6(g), 6(h) and 6(i). For noisy sequences of TownCentre, LatSVM and LatSVM-IN rank the third and fourth just behind HOG and C4. However, the high *Stability* values of HOG and C4 on TownCentre are meaningless, due to their low detection *Accuracy* on the reference video of TownCentre (less than 0.1).

It has been demonstrated that, the tail of singular values are dominated by the white Gaussian noise while the first few singular values are dominated by the image structure when performing SVD to noisy images [22], [25]. Likewise, the PCA-HOG used in LatSVM and LatSVM-IN can be a good way for increasing the detection accuracy on noisy video sequences, by only using the principle components which are more robust to white noise.

G. STABILITY WITH BRIGHTNESS VARIATION

The last row of Fig. 6 shows the *Accuracy* changing to brightness variation of the evaluated detectors. "0" on brightness level axis denotes the reference video sequences, while negative values indicate low-brightness sequences and positive values indicate high-brightness sequences.

From the *Stability* values of left-bottom corners in Fig. 7 and ranking in Fig. 6(j), 6(k) and 6(l), it can be seen that detectors with HOG based features alone (HOG, LatSVM and LatSVM-IN) are more stable with brightness variation comparing with other evaluated detectors, in line with the claim in [9] that HOG feature is invariant to illumination variation. Moreover, it can be seen that ACF and ACF-Cal which use LUV feature channel are more sensitive to brightness variation. Another interest finding is that, evaluated detectors shows higher stability with brightness variation than QP, resolution variation and additive white noise. One possibility can be that, brightness variation caused by environment illumination change has been widely studied by the community of pedestrian detection, and thus the brightness variation problem is better addressed in recent pedestrian detection studies.

V. SUMMARY AND DISCUSSION

In this paper, we have introduced the *DSurVD* for evaluating the robustness of pedestrian detectors to video distortions including H.264 compression distortion, resolution variation, additive white noise and brightness variation. Moreover, we give a thorough discussion of the detection robustness regarding to video quality degradation. The robustness is composed by the detection accuracy on good-quality reference videos A_{ref} and the performance stability on distorted videos. Based on the rate of accuracy degradation and monotonicity criteria, we define the detection stability mathematically. We also propose an intuitive robustness presentation method named *Robustness Quadrangle* which can be easily used to compare both the accuracies and stabilities between detectors. Usually, we treat the A_{ref} as the main attribute of robustness. However, when the A_{ref} of the compared detectors are close, as in many cases in practice, the stability measurement provides one more dimension to measure the robustness of the detectors.

With in-depth analysis of detection stability in Sec. IV, we have the following findings: 1) To H.264 compression distortion, Lbp feature can be an important complement to HOG feature in pedestrian detection. 2) Detectors trained with low model height performs more stable when the spatial resolution of the video reduces. However, training with unreasonable low model height may result in detection accuracy decrease, and cannot work well with high resolution cases. Obviously, more careful studies are called for generalization of the learnt models. 3) LatSVM shows more promising stability [16] to additive gaussian noise compared with other evaluated detectors. And this is possibly resulted by the PCA-HOG, which only consider the main structures of HOG features after PCA. 4) HOG shows the best stability with brightness variation among all the features used by the evaluated detectors.

Based on the robustness evaluation, the following two crucial cases are important to improve the detection robustness

in the future studies, distorted videos with quality under ''critical quality point'' and over-exposed videos with high brightness. To distorted videos with compression distortion, resolution reduction and additive white noise, the detection accuracy of most detectors does not gradually decrease when video quality drops, but decreases dramatically when the video quality reaches a critical point. Hence, efforts on extending the ''critical quality point'' could be one way to improve the detection stability. Compared with the pedestrian detection on low-brightness videos, detection on overexposed videos is an more challenging task and has been rarely studied.

As mentioned previously, much progress has been made to improve the detection accuracy in good-quality videos during the past decades. Comparing the evaluated pedestrian detectors in this study, the average miss rate on the high quality reference videos of *DSurVD* has been reduced from around 90% to 50%. However, quite less attention has been put on the research of the detection stability on surveillance videos with low quality. The in-depth analysis in this study have shown that there is still much room for improvement regarding to pedestrian detection in surveillance videos with low quality (often occurring in real-world situations).

REFERENCES

[1] *CAVIAR Dataset*. Accessed: Oct. 10, 2014. [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

[2] M. Andriluka, S. Roth, and B. Schiele, ''People-tracking-by-detection and people-detection-by-tracking,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[3] M. Andriluka, S. Roth, and B. Schiele, ''Monocular 3D pose estimation and tracking by detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 623–630.

[4] F. Bellard *et al.* (2014). *Ffmpeg*. [Online]. Available: http://www.ffmpeg.org

[5] R. Benenson, M. Omran, J. Hosang, and B. Schiele, ''Ten years of pedestrian detection, what have we learned?'' in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 613–627.

[6] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, ''Seeking the strongest rigid detector,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3666–3673.

[7] B. Benfold and I. Reid, ''Stable multi-target tracking in real-time surveillance video,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3457–3464.

[8] J. Cao, Y. Pang, and X. Li, ''Pedestrian detection inspired by appearance constancy and shape symmetry,'' *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5538–5551, Dec. 2016.

[9] N. Dalal and B. Triggs, ''Histograms of oriented gradients for human detection,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 886–893.

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona, ''Pedestrian detection: A benchmark,'' in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 304–311.

[11] P. Dollár, R. Appel, S. Belongie, and P. Perona, ''Fast feature pyramids for object detection,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[12] P. Dollár, C. Wojek, B. Schiele, and P. Perona, ''Pedestrian detection: An evaluation of the state of the art,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[13] M. Enzweiler and D. M. Gavrila, ''Monocular pedestrian detection: Survey and experiments,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[14] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, ''A mobile vision system for robust multi-person tracking,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, ''The Pascal visual object classes (VOC) challenge,'' *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, ''Object detection with discriminatively trained part-based models,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[17] J. Ferryman and A. Shahrokni, ''PETS2009: Dataset and challenge,'' in *Proc. 12th IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, Dec. 2009, pp. 1–6.

[18] K. Gu, G. T. Zhai, and M. Lin, ''The analysis of image contrast: From quality assessment to automatic enhancement,'' *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 284–297, Jan. 2015.

[19] E. Kafetzakis, C. Xilouris, M. A. Kourtis, M. Nieto, I. Jargalsaikhan, and S. Little, ''The impact of video transcoding parameters on event detection for surveillance systems,'' in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2013, pp. 333–338.

[20] P. Korshunov and W. T. Ooi, ''Video quality for face detection, recognition, and tracking,'' *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 7, no. 3, 2011, Art. no. 14.

[21] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami, and A. C. Kot, ''No-reference image blur assessment based on discrete orthogonal moments,'' *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 39–50, Jan. 2016.

[22] W. Liu and W. Lin, ''Additive white Gaussian noise level estimation in SVD domain for images,'' *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 872–883, Mar. 2013.

[23] S. Maji, A. C. Berg, and J. Malik, ''Classification using intersection kernel support vector machines is efficient,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[24] A. Milan, K. Schindler, and S. Roth, ''Detection- and trajectory-level exclusion in multiple object tracking,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3682–3689.

[25] M. Narwaria and W. Lin, ''SVD-based quality metric for image and video using machine learning,'' *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 347–364, Apr. 2012.

[26] T. Ojala, M. Pietikäinen, and D. Harwood, ''A comparative study of texture measures with classification based on featured distributions,'' *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.

[27] W. Ouyang, X. Zeng, and X. Wang, ''Learning mutual visibility relationship for pedestrian detection with a deep model,'' *Int. J. Comput. Vis.*, vol. 120, no. 1, pp. 14–27, 2017.

[28] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, ''Pedestrian detection with spatially pooled features and structured ensemble learning,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1243–1257, Jun. 2016.

[29] P. Peng, Y. Tian, Y. Wang, J. Li, and T. Huang, ''Robust multiple cameras pedestrian detection with multi-view Bayesian network,'' *Pattern Recognit.*, vol. 48, no. 5, pp. 1760–1772, 2015.

[30] N. Ponomarenko *et al.*, ''TID2008—A database for evaluation of full-reference visual quality assessment metrics,'' *Adv. Mod. Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[31] *IEEE Standard Glossary of Software Engineering Terminology*, IEEE Standard 610.12-1990, 1990.

[32] H. Schwarz, D. Marpe, and T. Wiegand, ''Overview of the scalable video coding extension of the H.264/AVC standard,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[33] J. Shen, X. Zuo, J. Li, W. Yang, and H. Ling, ''A novel pixel neighborhood differential statistic feature for pedestrian and face detection,'' *Pattern Recognit.*, vol. 63, pp. 127–138, Mar. 2017.

[34] G. Shu, A. Dehghan, and M. Shah, ''Improving an object detector and extracting regions using superpixels,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3721–3727.

[35] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele, ''Learning people detectors for tracking in crowded scenes,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1049–1056.

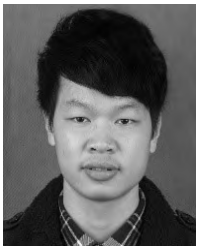
[36] Y. Tian, P. Luo, X. Wang, and X. Tang, ''Pedestrian detection aided by deep learning semantic tasks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5079–5087.

[37] X. Wang, T. X. Han, and S. Yan, ''An HOG-LBP human detector with partial occlusion handling,'' in *Proc. Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 32–39.

[38] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, ''Overview of the H.264/AVC video coding standard,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[39] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 794–801.
[40] J. Wu, C. Geyer, and J. M. Rehg, "Real-time human detection using contour cues," in *Proc. Int. Conf. Robot. Automat.*, May 2011, pp. 860–867.
[41] J. Wu and J. M. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug. 2011.
[42] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Multi-pedestrian detection in crowded scenes: A global view," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3124–3129.
[43] Y. Yuan, W. Lin, and Y. Fang, "Is pedestrian detection robust for surveillance?" in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 2776–2780.
[44] Y. Yuan, S. Emmanuel, Y. Fang, and W. Lin, "Visual object tracking based on backward model validation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 11, pp. 1898–1910, Nov. 2014.
[45] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1259–1267.
[46] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed Haar-like features improve pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 947–954.

YUMING FANG (SM'17) received the B.E. degree from Sichuan University, Chengdu, China, the M.S. degree from the Beijing University of Technology, Beijing, China, and the Ph.D. degree from Nanyang Technological University, Singapore. He is currently a Professor with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. He has authored and co-authored over 90 academic papers in international journals and conferences in the areas of multimedia processing. His research interests include visual attention modeling, visual quality assessment, image retargeting, computer vision, 3-D image/video processing, and so on. He serves as an Associate Editor for the IEEE Access, and he is on the Editorial Board of *Signal Processing: Image Communication*.



GUANQUN DING is currently a Graduate Student with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include saliency detection, object detection, computer vision, and so on.

YUAN YUAN received the B.E. degree in electronic engineering from the Beijing University of Post and Telecommunication, Beijing, China, in 2011. He is currently pursuing the Ph.D. degree with the School of Computer Engineering, Nanyang Technological University, Singapore. His current research interests include visual surveillance, object tracking, and visual attention.

WEISI LIN (M'92–SM'98–F'16) received the Ph.D. degree from King's College London, U.K. He served as the Lab Head of visual processing, Institute for Infocomm Research, Singapore. He is currently an Associate Professor with the School of Computer Engineering. His technical expertise includes perceptual modeling and evaluation of multimedia signals, image processing, and video compression, in which he has published 160 journal papers and 230 conference papers, filed seven patents, authored two books, edited three books, and written nine book chapters. He is a Chartered Engineer, a fellow of IET, and an Honorary Fellow of the Singapore Institute of Engineering Technologists. He has been a Technical Program Chair of IEEE ICME 2013, PCM 2012, QoMEX 2014, and VCIP 2017. He chaired the IEEE MMTC Special Interest Group on QoE (2012–2014). He has been a keynote/invited/panelist/tutorial speaker in over 20 international conferences, as well as a Distinguished Lecturer of the Asia-Pacific Signal and Information Processing Association from 2012 to 2013 and the IEEE Circuits and Systems Society from 2016 to 2017. He served as a guest editor for seven special issues in different scholarly journals. He was an Associate Editor (AE) of the IEEE Transactions on Multimedia and the IEEE Signal Processing Letters. He is an AE of the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, and the *Journal of Visual Communication and Image Representation*.

HAIWEN LIU (M'04–SM'13) received the B.S. degree in electronic system and the M.S. degree in radio physics from Wuhan University, Wuhan, China, in 1997 and 2000, respectively, and the Ph.D. degree in microwave engineering from Shanghai Jiao Tong University, Shanghai, China, in 2004. From 2004 to 2006, he was a Research Assistant Professor with Waseda University, Kitakyushu, Japan. From 2006 to 2007, he was a Research Fellow with Kiel University, Kiel, Germany, where he received the Alexander von Humboldt Research Fellowship. From 2007 to 2008, he was a Professor with the Institute of Optics and Electronics, Chengdu, China, where he was supported by the 100 Talents Program of Chinese Academy of Sciences, Beijing, China. From 2009 to 2017, he was a Chair Professor with East China Jiaotong University, Nanchang, China. Since 2017, he has been a Chair Professor with Xi'an Jiaotong University, Xi'an, China. He has authored over 100 papers in international and domestic journals and conferences. He has served as a technical program committee member for many international conferences. He is an Associate Editor of the *International Journal of RF and Microwave Computer-Aided Engineering* and a Leading Guest Editor of the *International Journal of Antennas and Propagation*.