International Journal of Multimedia Data Engineering and Management
Volume 10 • Issue 1 • January-March 2019

A No-Reference Image Quality Model for Object Detection on Embedded Cameras

Lingchao Kong, University of Cincinnati, Cincinnati, USA
Ademola Ikusan, University of Cincinnati, Cincinnati, USA
Rui Dai, University of Cincinnati, Cincinnati, USA
Jingyi Zhu, University of Cincinnati, Cincinnati, USA
Dara Ros, University of Cincinnati, Cincinnati, USA

ABSTRACT

Automatic video analysis tools are an indispensable component in imaging applications. Object detection, the first and the most important step for automatic video analysis, is implemented in many embedded cameras. The accuracy of object detection relies on the quality of images that are processed. This paper proposes a new image quality model for predicting the performance of object detection on embedded cameras. A video data set is constructed that considers different factors for quality degradation in the imaging process, such as reduced resolution, noise, and blur. The performances of commonly used low-complexity object detection algorithms are obtained for the data set. A no-reference regression model based on a bagging ensemble of regression trees is built to predict the accuracy of object detection using observable features in an image. Experimental results show that the proposed model provides more accurate predictions of image quality for object detection than commonly known image quality measures.

KEYWORDS

Automatic Video Analysis, Image Distortion, Image Quality Assessment, No-Reference, Object Detection

DOI: 10.4018/IJMDEM.2019010102

INTRODUCTION[1]

Wireless embedded camera sensors have become ubiquitous components in various imaging applications, such as public safety and security systems, smart building operations, intelligent transportation, and remote health care. Rather than merely presenting raw data collected by camera sensors to the user, an application usually aims to automatically discover and extract meaningful information from the camera sensors and to achieve as much autonomy as possible in the physical system. Automatic video data analysis tools, which could detect, recognize, track objects of interest, and understand their behaviors, have become indispensable components in today's imaging applications.

The performance of automatic analysis methods relies on the quality of images that are processed. It is therefore essential to introduce objective metrics for predicting the quality of images evaluated by automatic analysis algorithms. In the field of image quality assessment (IQA), a diverse range of image quality models, ranging from full-reference to reduced-reference and no-reference ones, were designed for predicting the perceptual quality evaluated by human subjects (Ma et al., 2018, pp. 1202-1213; Wang et al., 2018, pp. 1-14; Wang, Bovik, Sheikh, & Simoncelli, 2004, pp. 600-612).

The quality of a video sequence judged by an automatic analysis algorithm, however, is not necessarily sensitive to the same factors that drive human perceptions. The perceptual image quality assessments usually try to emulate known characteristics of the human visual system (HVS), such as the contrast sensitivity and the visual attention mechanisms. The contrast sensitivity mechanism means that the HVS is sensitive to the relative luminance change rather than the absolute luminance change (Wang et al., 2004, pp. 600-612). The visual attention mechanism is that only a local area in the image can be perceived with high resolution by the human observer at one time instance at typical viewing distances, due to the foveation feature of the HVS (Yang et al., 2016, pp. 3475-3488). On the other hand, automatic analysis methods run by machines can "perceive" the absolute luminance change precisely and have a better global "view". For example, the problem of evaluating motion imagery quality for tracking in airborne reconnaissance systems was studied in Irvine and Wood's research (2013, p. 87130Z). It was found that automated target detection algorithms are less sensitive to spatial resolution than humans, but factors such as jitter in the temporal domain, texture complexity, edge sharpness, and level of noise have a strong effect on the performance of target detection. In our recent work (Kong, Dai, & Zhang, 2016, pp. 3797-3801), we found that unlike human beings who can easily extract and focus on a moving object from a blurred background, the performance of object detection algorithms can be affected by the quality of the background. These results suggest that new models are needed for evaluating the quality of images from the perspective of automatic analysis algorithms.

In a wireless imaging system, automatic analysis could be deployed using two strategies: in the central server on compressed videos; or at the local cameras on uncompressed videos as a preprocessing step. The impact of video compression on the accuracy of analysis algorithms has been studied in some recent works (Tahboub, Reibman, & Delp, 2017, pp. 4192-4196; Zhong, & Reibman, 2018, pp. 1-6), which aim at finding the optimal compression rates under a quality requirement. Apart from the distortion introduced by compression, the quality of an image or a video could be degraded during the data acquisition or sensing process, e.g., distortion caused by noise or motion blur, or reduced image resolution due to storage or bandwidth constraints on embedded cameras. These factors should also be taken into consideration to evaluate the quality of an image.

Object detection is the first and the most important step in the process of automatic analysis, because the detected objects provide a focus of attention for the following tasks such as tracking and recognition. In this paper, we propose a blind regression model based on a bagging ensemble of trees to predict the performance of object detection on an image. The model utilizes local features in an image such as edge and oriented gradient and global features including image gradient and estimated object size, which could be easily extracted from an image. The model is trained using a large number of images with different scene characteristics and four types of distortions including

noise, Gaussian blur, motion blur, and reduced spatial resolution. The accuracy of the proposed model is evaluated through extensive experiments on a separate test data set and compared with commonly used full-reference and popular no-reference IQA measures. This article extends our recent conference publication (Kong, Ikusan, Dai, & Zhu, 2019, pp. 216-221) with more comprehensive experimental results and discussions.

RELATED WORK

There are only a few studies on the problem of quality evaluation for automatic analysis algorithms. Image quality assessment for face recognition applications was studied in researches by Abaza, Harrison, and Bourlai (2012, pp. 3103-3170); Gunasekar, Ghosh, and Bovik (2014, pp. 2119-2131); Pulecio, Benítez-Restrepo, and Bovik (2017, pp. 805-809), Five quality factors were evaluated, including contrast, brightness, focus, sharpness, and illumination, and a face image quality index combining the five factors was proposed in a research by Abaza et al. (2012, pp. 3103-3170). In Pulecio et al. (2017, pp. 805-809), natural scene statistics was used to detect degradation of infrared images for face recognition. In Gunasekar et al. (2014, pp.2119-2131), the degradation in the performance of face detectors were quantified considering different factors including noise, blur, and compression.

There are also a few studies on the quality for target detection, target tracking, and event detection for airborne reconnaissance applications. In Irvine and Nelson (2009, p. 73350L), the applicability of the National Imagery Interpretability Ratings Scale (NIIRS) to an automated target detection algorithm was examined, and it was found that NIIRS is not a good predictor of target detection performance. In reseaches by Irvine, Wood, Reed, and Lepanto (2013, pp. 1-9) and Irvine and Wood (2013, p. 87130Z), the impacts of video frame rate and two spatial factors (noise and spatial resolution) on the tracker performance were investigated.

The aforementioned studies investigated the performance of automatic analysis on specific applications like face recognition and airborne reconnaissance. Our work advances the state of the art by addressing the challenge of building a more general quality prediction model for a wide range of object detection algorithms and diverse scene characteristics. Moreover, our model considers four common types of distortions during the imaging process.

DATA SET AND OBJECT DETECTION MEASURE

We have selected 10 high resolution original video sequences with different scene characteristics, illumination levels, and object scales. Among them, 5 videos are chosen from the Multiple Object Tracking (MOT) dataset (Milan, Leal-Taixé, Reid, Roth, & Schindler, 2016), and 5 videos are chosen from the Duke Multi-Target Multi-Camera Tracking (DM) dataset (Ristani, Solera, Zou, Cucchiara, & Tomasi, 2016, pp. 17-35). The resolutions of these videos are mostly 1920 × 1080 except for one

Figure 1. Snapshots of video data set

video with 640 × 480 resolution, and the average number of frames is 741. The snapshot of these videos is shown in Figure 1.

To understand how the performance of object detection could be affected by image distortions, we have generated different distorted video sequences based on the original videos, where the distortion falls into four types: Gaussian blur, motion blur, imaging noise, and reduced spatial resolution. For each type of distortion, distortion levels are set to low level and high level, and the simulation parameters and setting are selected based on the experiments conducted in recent works (Ding et al., 2018, pp. 1002-1014; Ma et al., 2016, pp. 1004-1016). The blurring effect of a video is generated by 2D circularly symmetric Gaussian blur kernels with standard deviations of 1.2 and 6.5 for low level and high level, respectively. The motion blur is simulated to approximate the linear motion of a camera by 5 and 20 pixels with an angle of 45 degrees for low and high levels, respectively. White Gaussian noise is added to the original images, where variances are set to be 0.001 and 0.022 for low and high levels, respectively. For reduced spatial resolution, 1:2 and 1:4 down-sampling rates are applied in both horizontal and vertical directions on the original images. The related simulation parameters and setting are summarized in Table 1. Samples for the four types of distortion are shown in Figure 2, in which the original image frame is the 581_{th} frame of DMcam01 video. Figure 2 (a) shows the image with reduced spatial resolutions, which includes 3 resolutions overlaying in one image, corresponding to original, half, and quarter resolutions in both horizontal and vertical directions. Figure 2 (b) is a sample of blur to simulate out-of-focus blur, Figure 2 (c) is a sample of motion blur to simulate camera shake during exposure, and Figure 2 (d) is a sample of white noise to simulate imaging noise in low-light scenarios. For each original video sequence shown in Figure 1, we have generated a total number of 8 distorted videos, including 2 videos from each type of distortion. This results in a total number of 90 video sequences (including the original ones) in our data set.

There are two categories of object detection algorithms in the field of computer vision: one based on building models of backgrounds and the other based on building models for objects. Algorithms based on background modeling require multiple frames to build a stable background, while methods based on object modeling could generate detection results on a single image. In this work, we aim at predicting the quality of single images in a fast manner, such that the wireless embedded imaging system could adjust its sensing strategy based on the predicted quality and energy supply. Therefore, we focus on the performance of low-complexity object modeling methods. We use the following three representative lightweight algorithms based on object modeling:

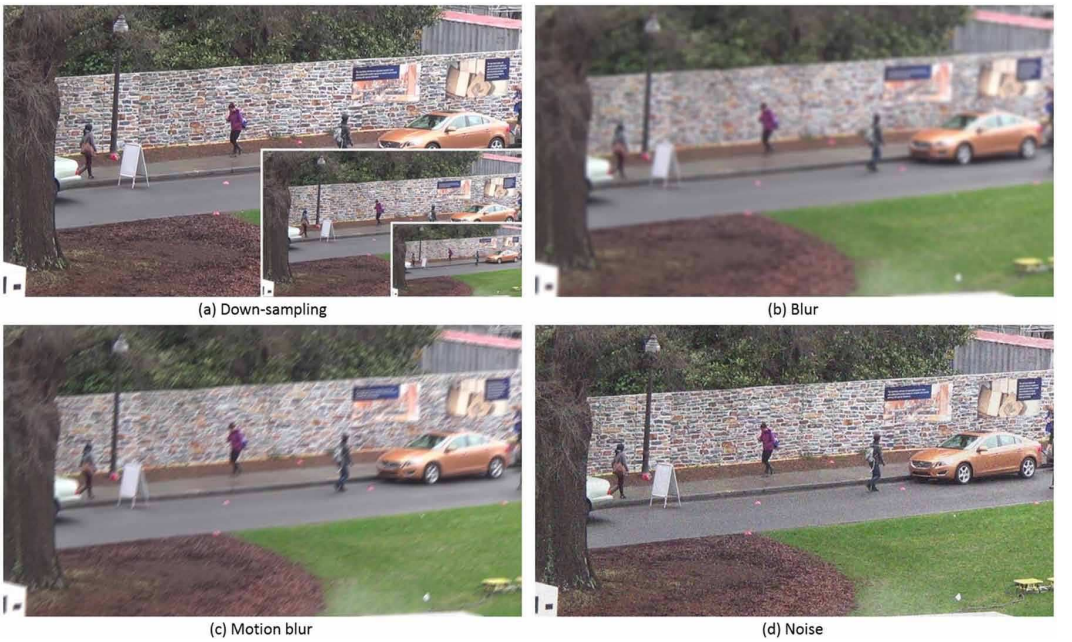
1. Histograms of Oriented Gradients (HOG) (Dalal & Triggs, 2005, pp. 886-893)
2. Discriminatively Part Models (DPM) (Felzenszwalb, Girshick, McAllester, & Ramanan, 2010, pp. 1627-1645)
3. Locally Decorrelated Channel Features (LDCF) (Nam, Dollár, & Han, 2014, pp. 424-432)

The evaluation measures for object detection could be either sequence-based or image-based. Since our goal is to predict the performance of object detection once an image is taken, we evaluate the object detection accuracy of each frame in a video. The Frame Detection Accuracy (FDA) is a comprehensive metric that accounts for important measures of system performance (such as number of objects detected, missed objects, false positives, and localization error of detected objects) in a single score (Kasturi et al., 2009, pp. 319-336). For a given frame, the optimal matching pairs is assigned firstly by computing the spatial overlap between ground truth and detected objects. Then, the FDA measure calculates the spatial overlap between the ground truth and system output objects as a ratio of the spatial intersection between the two objects and the spatial union of them. The sum of all of the overlaps was normalized over the average number of ground truth and detected objects. For one image, where there are N_G ground-truth objects G and N_D detected objects D, N_m is the number of mapped object pairs, FDA is defined as

Table 1. Distortion parameters

Distortion	Reduced spatial resolution	Gaussian blur	Motion blur	Imaging noise
	Down-sampling resolution	2D circularly symmetric kernel (standard deviation)	Linear motion with an angle of 45 degrees (pixels)	White Gaussian noise (variance)
Low level	1:2	1.2	5	0.001
High level	1:4	6.5	20	0.022

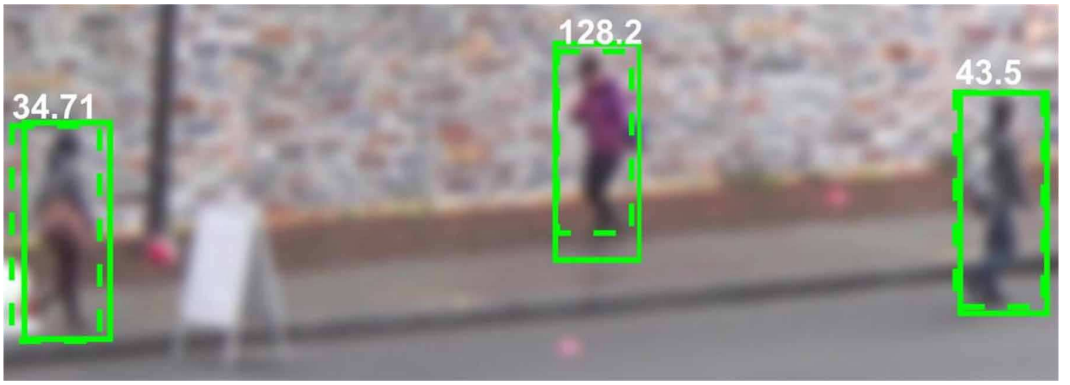
Figure 2. Samples of different distortions



$$FDA = \frac{\sum_{i=1}^{N_m} \frac{G_i \cap D_i}{G_i \cup D_i}}{\left(N_G + N_D\right)/2} \tag{1}$$

A detection system needs to take an image and return a bounding box and a confidence for each detection. The provision of a confidence level allows results to be ranked such that the trade-off between false positives and false negatives can be evaluated, without defining arbitrary costs on each type of classification error (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010, pp. 303-338). However, the original FDA measure does not reflect the trade-off between false positives and false negatives. Thus, we introduce a revised FDA measure, rFDA for short, which is the average of FDA based on different thresholds (T) of detection confidence levels (C). $rFDA$ is defined as,

Figure 3. Detection sample



$$rFDA = \sum_{j=1}^{N_m} \left(\frac{\sum_{i=1}^{N_{T_j}} \frac{G_i \cap D_i}{G_i \cup D_i}}{\frac{N_G + N_D}{2}} \right) / N_m, \tag{2}$$

where N_m is the number of mapped object pairs, N_{T_j} is the number of true positives when the threshold of detection confidence T_j equals to C_j, $j \in \{1, ..., N_m\}$, and C_j denotes the detection confidence level of the *j*-th mapped detected object.

The original FDA measure in can be regarded as $FDA_{T(min)}$, which uses the minimum detection confidence level $C(min)$ in mapped pairs as threshold such that all mapped object pairs are true positives. A detection sample is shown in Figure 3, which corresponds to a part of the 581_{th} frame of high blur distorted DMcam01 video by LDCF detector. The ground truth is highlighted in solid line, and three detected objects in dash line with confidence levels 34.71, 128.2, and 43.5, respectively. When the threshold T equals to the minimum confidence $C(min)$, i.e., $T = 34.71$, three detection results are all true positives, which is the same with the original FDA definition; when $T = 43.5$, only two detection results are regards as true positive. On the other hand, the SSIM (0.51) and PSNR (21.14 dB) values of this image actually are quite low and poor, however, detection performance is pretty good, which indicates that the popular image quality assessments can not reflect the detection quality.

In order to validate the proposed rFDA measure, we compare the correlation of FDA and rFDA with Average Miss Rate (AMR), which is the most popular metric used in the object detection area. The AMR of an image sequence (Dollar, Wojek, Schiele, and Perona, 2012, pp. 743-761) can be determined as follows: first, a detected object and a ground truth form a match if they overlap sufficiently, which is evaluated by the ratio of the intersection between two objects and the union of them, and a threshold ratio of 0.5 is commonly used; then, the miss rates against false positives per image (FPPI)is plotted (using log-log plots) by varying the threshold on detection confidence; finally, the log-average miss rate is used to summarize the detector performance by averaging miss rate at nine FPPI rates evenly spaced in log-space in the range of 0.01 to 1. Since AMR is calculated based on the entire image sequence, we measure the detection performance for the whole sequence using Sequence Frame Detection Accuracy (SFDA) introduced by Kasturi et al. (2009, p. 325). SFDA is an

average of the FDA measured over all frames in sequence. The average is normalized to the number of frames in the sequence where at least a ground truth or a detected object exists. SFDA is formulated as

$$SFDA = \frac{\sum_{t=1}^{N_{frames}} FDA\left(t\right)}{\sum_{t=1}^{N_{frames}} \exists\left(N_G^t \, ORN_D^t\right)}, \tag{3}$$

where N_G^t and N_D^t denote the number of ground-truth objects and the number of detected objects in frame t, respectively, N_{frames} is the number of frames in the sequence, and $FDA\left(t\right)$ is the FDA value for the single frame t.

We compare the usages of the original FDA measure and the proposed rFDA measure in the computation of sequence-level FDA, i.e., SFDA and SrFDA. The correlation coefficients of SFDA with AMR and the ones of SrFDA with AMR for three different detectors are summarised on Table 2. We can find that there are obvious improvements comparing SrFDA's correlation with SFDA's for all the three detectors. Specifically, gains are 0.0426, 0.0936, and 0.0365 for DPM, HOG, and LDCF detectors, respectively. The average of correlation coefficients between SrFDA and AMR for the three detectors reaches 0.9095, and the average gain is 0.0576, which indicates that SrFDA is more consistent with AMR. It indicates that rFDA can depict the performance of object detection and it is a better metric for single image detection performance. In addition, the variations of each frame's FDA and rFDA in the image sequences are inspected based on standard deviation, and the results are summarized in Table 2. We can notice that the variations of rFDA are always smaller than the ones of FDA for the three different detectors, which indicates that rFDA can reduce arbitrary fluctuations and maintain more stable measurements.

The correlation between the different object detectors are also investigated, as shown in Table 3. The correlation coefficients are all above 0.83, and the average of them reaches 0.8501. Although the operating principles of the three detectors are different, the correlation results indicate that their detection performances are consistent. Therefore, we target at predicting the average performance of the three detectors in the proposed quality model.

We have visualized the distribution of rFDA values for the images in our data set. Figure 4 compares the rFDAs obtained from the original images and the images with high and low distortions. Figure 4 (a), (b), (c), and (d) correspond to the results on down-sampled, blurred, motion blurred, and noisy images, respectively. The "ori" label denotes the rFDAs of the original images, the "ds2" and "ds4" labels denote the ones of 1:2 and 1:4 down-sampling version, and the pairs of "lb" and "hb", "lm" and"hm", "ln" and"hn" denote the ones of low level and high level for blur, motion blur, and noise, respectively. We can find that rFDAs have different ranges of values when different levels of distortions are introduced for all the four types of distortions.

Table 2. Comparison of rFDA and FDA in correlation and variation

Detector		**DPM**	**HOG**	**LDCF**	**Average**	**Δ**
Correlation coefficients	*SFDA*	0.9326	0.7134	0.9097	0.8519	-
	SrFDA	0.9752	0.8070	0.9462	0.9095	0.0576
Variation (standard deviation)	*FDA*	0.2393	0.1506	0.2873	0.2257	-
	rFDA	0.1731	0.1020	0.1898	0.1550	-0.0707

Table 3. The consistency between different detectors

DPM vs. HOG	DPM vs. LDCF	HOG vs. LDCF	Average
0.8641	0.8471	0.8390	0.8501

BLIND MODEL FOR PREDICTION OF RFDA

In this section, we introduce 13 efficient local and global features and a supervised learning algorithm to build a regression model for rFDA.

Boundary information in an image plays an important role in object detection and pattern recognition since boundaries represent the transition regions between objects and background where the image intensities vary abruptly or have discontinuities. Gradient is a good indicator for the variance of image intensities. For an image $f\left(x,y\right)$, the gradient of f at location $\left(x,y\right)$ is defined as the two dimensional column vector: $\left[\frac{\partial f}{\partial x},\frac{\partial f}{\partial y}\right]^T$,where $\frac{\partial f}{\partial x}=f\left(x+1,y\right)-f\left(x-1,y\right)$, and $\frac{\partial f}{\partial y}=f\left(x,y+1\right)-f\left(x,y-1\right)$ using finite difference filters. The magnitude and direction of this gradient at location $\left(x,y\right)$ are given by

$$mag\left(x\right)=\sqrt{\left(\frac{\partial f}{\partial x}\right)^2+\left(\frac{\partial f}{\partial y}\right)^2}, \tag{4}$$

$$dir\left(x,y\right)=tan^{-1}\left[\frac{\partial f/\partial y}{\partial f/\partial x}\right]. \tag{5}$$

One sample of image gradient is shown in Figure 5. The original image in Figure 5 (a) is a portion of the 581_{th} frame of DMcam01 video, and Figure 5(b) and (c) show the corresponding image gradient directions and magnitudes, respectively. We can observe that the image gradient direction and magnitude can depict the boundary of objects precisely. Thus, the statistical properties of gradient could be used to depict the characteristics of an image. We calculate 4 related features: (1) meanGmag: the average of gradient magnitude; (2) stdGmag: the standard deviation of gradient magnitude; (3) meanGdir: the average of gradient direction; (4) stdGdir: the standard deviation of gradient direction.

Figure 4. Distribution of rFDA on images with different types of distortions

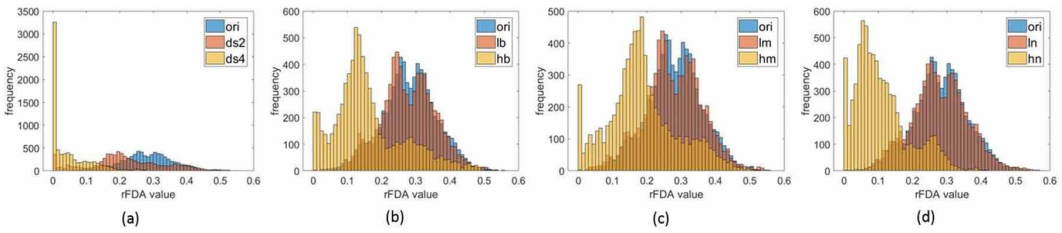
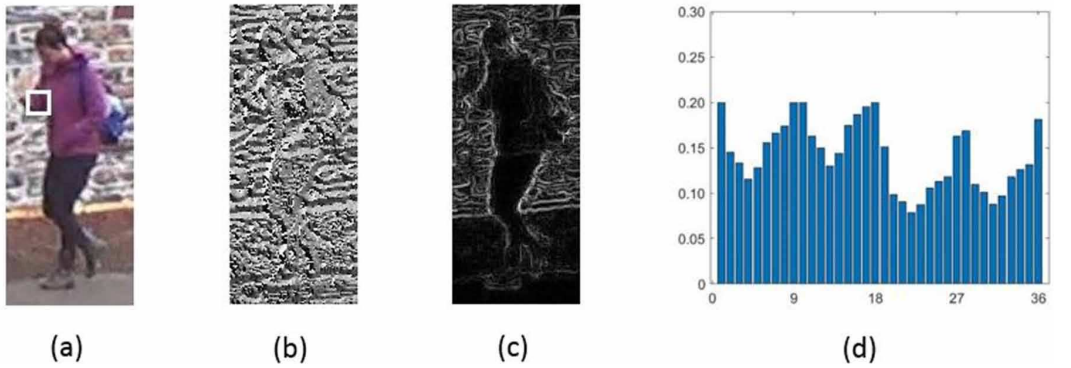


Figure 5. Sample of image gradient and HOG descriptor



The local oriented gradient can describe object appearance and shape through counting occurrences of gradient orientation in localized portions of an image based on the HOG descriptor defined in a work of Dalal and Triggs (2005, pp. 886-893). The local window for one HOG descriptor is set as 16 × 16 pixels, and the number of orientation bins for one HOG descriptor is set as 9. For each local window, a histogram of gradients, with each gradient quantized by its angle and weighed by its magnitude, is calculated. The gradient of color images is computed separately for each color channel and the one with maximum magnitude is selected. For each histogram with 9 orientation bins,4 different normalizations using adjacent histograms are employed, which results in a 36-dimensional feature vector. One sample of HOG descriptor is shown in Figure 5 (d), where the location of the local window is marked with a white square in Figure 5 (a). We can observe that the trend/shape is indeed repeated 4 times for every 9 bins due to 4 different normalizations. The average frequency w_m and the frequency's variation level w_s of the histogram's bins, are defined to one window as follows:

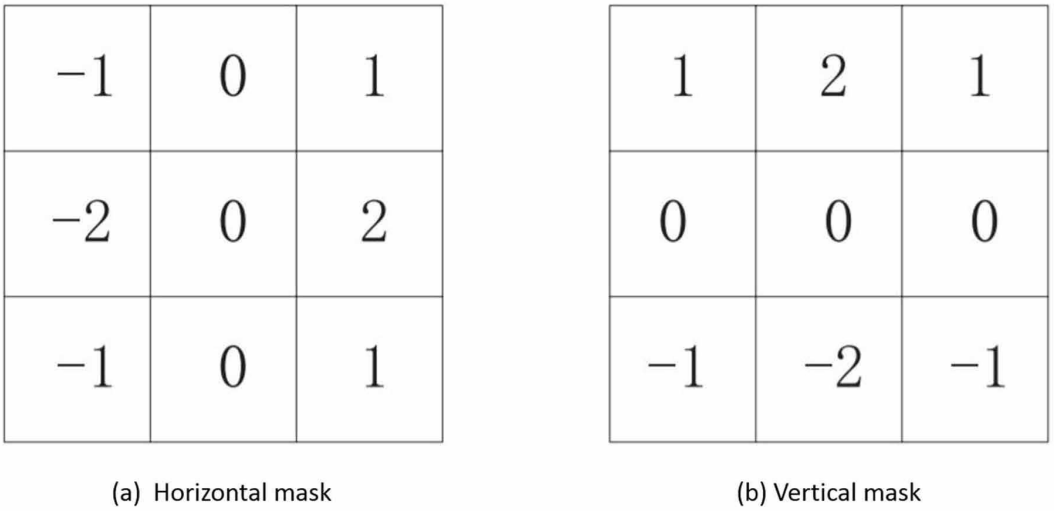
$$w_m = \sum_{i=1}^{N_b} h_i / N_b, \tag{6}$$

$$w_s = \sqrt{\sum_{i=1}^{N_b} (h_i - w_m)^2 / \left(N_b - 1\right)}, \tag{7}$$

where h_i is the frequency of the i_{th} bin in a local window, and N_b is the number of bins in a local window. Based on two statistical values for one local window, 4 related features are calculated: (5) hog_mm: the average of every blocks' w_m ; (6) hog_ms: the standard deviation of every blocks' w_m ; (7) hog_sm: the average of every blocks' w_s ; (8) hog_ss: the standard deviation of every blocks' w_s .

The boundary or edge, representing transition areas between objects and background, is obtained by Sobel operator through convolving the image with two 3×3 kernels in the horizontal and vertical directions. The convolution masks for horizontal and vertical directions are shown in Figure 6.The local

Figure 6. Sobel masks



information of edge is collected based on a block of 16 × 16 pixels, 4 related features are calculated: (9) edge_mm: the average of every blocks' average; (10) edge_ms: the standard deviation of every blocks' average; (11) edge_sm: the average of every blocks' standard deviation; (12) edge_ss: the standard deviation of every blocks' standard deviation.

If the size of an object is too small or too large in the image, it is hard to detect the object from the background. The last feature is designed as: (13) estimated object size, which is calculated approximately using the method proposed by Yang, Li, Li, and Li (2016, pp. 3475-3488). First, a contour-based spatial prior is extracted based on the layout of edges in the given scene along a fast non-selective pathway, which provides a rough, task-irrelevant, and robust estimation of the locations where the potential objects are present. The contour-represented layout in the non-selective pathway is used as the initial guidance to estimate the locations and sizes of objects and the relative importance of low-level local cues. Then, local features such as color, luminance, and texture, are extracted in parallel along the selective pathway. Finally, Bayesian inference is used to auto-weight and integrate the local cues guided by contour-based spatial prior and to predict the exact locations of objects. The objects are further enhanced via iterative processing to refine the prior guidance as the final prediction.

We use the bootstrap aggregating, or bagging, ensemble of trees to train a regression model to predict detection performance based on the extracted 13 features on a single image (Breiman, 2001, pp. 5-32). Every decision tree in the bagging ensemble is grown on an independently drawn bootstrap replica of input observations. The ensemble tree prediction is formed by taking the average over base learners. The tuning parameters of ensemble trees include the number of trees and minimum leaf size to control the tree depth.

PERFORMANCE EVALUATION

To evaluate the performance of the proposed model, we divide the entire data set into a training set and testing set, which are described in Table 4. The total number of images in our data set is 66672. The images from 8 raw videos and their distorted versions are used for training (75.03%), and the images from the remaining 2 raw videos and their distorted versions are used for testing. Through 5-fold cross validation during the training procedure, 30 base learners and a minimum leaf size of 8 are used to build the ensemble of trees.

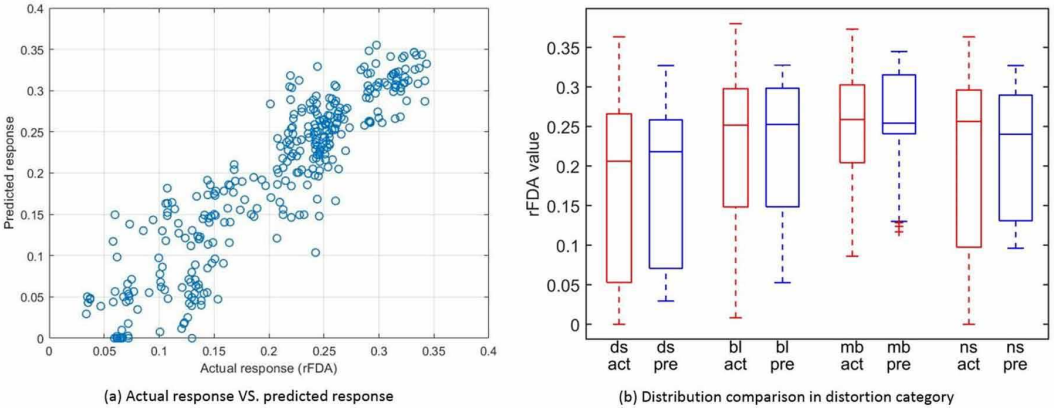
Table 4. Learning setting

Category	Video name	Image Number	Percentage
Training set	MOT17-02, MOT17-10, MOT15-02, DMcam01, DMcam02, MOT17-13, DMcam04, DMcam08	50022	75.03%
Testing set	MOT17-04, DMcam06	16650	24.97%

First, the regression performance of the proposed model is investigated. Figure 7 (a) shows the scatter figure of the actual response VS. the predicted response. There are a huge number of observations (16650 images) in the testing data set, and one point is selected from every 50 observations to generate a clear figure. The perfect regression results should be all on the diagonal line, and most of the predictions in our proposed model are near or on the diagonal line, which indicates that the regression of proposed model can depict the image quality for object detection quite well. Figure 7 (b) illustrates the distributions of the actual response and the predicted response in four distortion categories, down-sampling in the spatial domain (ds), blur (bl), motion blur (mb), and imaging noise (ns), in which the actual responses (act) are in red color and the predicted responses (pre) are in blue color with wider boxes. We can find that the 25_{th} and 75_{th} quartiles and the medians of the predicted responses are all close to the actual responses in the distribution of four distortion categories, indicating that the proposed model can accurately predict image quality for object detection for different types of distortions.

The regression performance of proposed model on the testing data set is measured in terms of Root Mean Square Error (RMSE), R^2, $adjR^2$. Mean Squared Error (MSE), and Mean Absolute Error (MAE), as shown in Table 5. Among these metrics, smaller values of RMSE, MSE and MAE indicate better performance. R^2, or coefficient of determination, is always smaller than 1 and usually larger than 0. Adjusted R^2, short for $adjR^2$, adjusts R^2 for the number of explanatory terms (features) in a model relative to the number of observations. R^2 and $adjR^2$ values close to 1 indicates good regression performance. From Table 5, we can find that the overall performance in terms of RMSE, MSE and MAE is all quite close to 0, and both R^2 and $adjR^2$ reaches 0.814, which indicates that the proposed model fits data well and that only a few features can explain the observations. The performances of specific distortion categories are also inspected and summarized in Table 5. For the down-sampling and the blur categories, the values of RMSE, MSE, and MAE are all less than the

Figure 7. Regression performance



(a) Actual response VS. predicted response

(b) Distribution comparison in distortion category

Table 5. Regression metrics

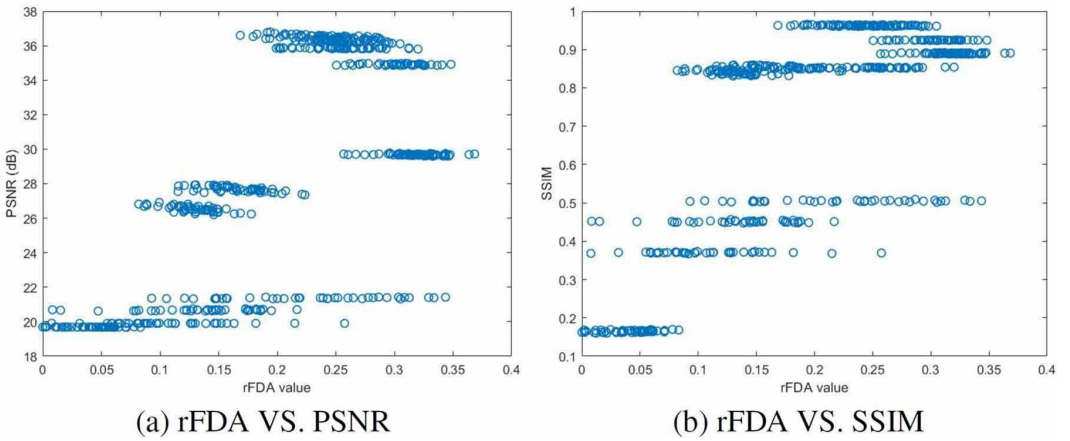
Metrics		RMSE	R^2	$adjR^2$	MSE	MAE
Overall performance		0.0428	0.8147	0.8146	0.0018	0.0324
Different distortion category performance	Down-sampling	0.0415	0.8116	0.8110	0.0017	0.0339
	Blur	**0.0334**	**0.8321**	**0.8316**	**0.0011**	**0.0250**
	Motion blur	0.0400	0.6208	0.6196	0.0016	0.0293
	Noise	0.0589	0.7204	0.7196	0.0035	0.0479

ones of overall performance, and the values of R^2 and adj$R^2$are similar or even larger than the ones of overall performance. For the motion blur category, the values of RMSE, MSE, and MAE are all smaller than the the ones of overall performance except for slightly lower values of R^2 and adjR^2. For the noise category, the values of RMSE, MSE, and MAE is close to the ones of overall performance, and the values of R^2 and adjR^2 are higher than the ones in the motion blur category. Generally speaking, the proposed model can handle different distortion categories and achieve a decent overall performance.

The performance of the proposed model is also compared with commonly used full-reference IQAs. Although the proposed model is a blind, or no-reference, image quality estimator for object detection, two full-reference IQAs, i.e. PSNR and SSIM, are compared in terms of the Linear Correlation Coefficient (LCC), the Spearman Rank Order Correlation Coefficient (SROCC), and the Kendall Rank Correlation Coefficient (KRCC). Because full-reference PSNR and SSIM measures could not evaluate the quality of down sampling versions and original video sequences, results from these images are excluded in this comparison. The correlation results are shown in Table 6. The correlation coefficients of LCC and SROCC for the proposed model reach above 0.90, while the ones for SSIM and PSNR are all below 0.50; the correlation coefficients of KRCC for the proposed model also reach above 0.70, which is more than twice over the ones for SSIM and PSNR. The results show that the proposed model is a good predictor for the image quality for object detection, and SSIM and PSNR cannot be good indicators for the image quality for object detection. The conclusion also can be drawn from Figure 8, in which scatter figures between PSNR, SSIM and rFDA values are plotted. From Figure 8, we can find that there is no significant relationship between either PSNR or SSIM and rFDA values. The reason is that SSIM and PSNR are designed for the perceptual quality but not for the quality evaluated by object detection algorithms.

Then, comparison is conducted between the proposed model and two popular no-reference IQAs, BRISQUE and BLIINDS-II, in terms of the LCC, the SROCC, and the KRCC. BRISQUE (Mittal, Krishna, and Bovik, 2012, pp. 4695-4708) is a distortion-generic no-reference IQA model, which exploits scene statistics of locally normalized luminance coefficients in spatial domain to quantify possible losses of "naturalness" in the image. BLIINDS-II (Saad, Bovik, and Charrier, 2012, pp. 3339-3352) is a blind IQA algorithm using a Bayesian inference approach on extracted features that are based on a natural scene statistics model of discrete cosine transformation coefficients. All images in the testing set are included in this comparison thanks to the no-reference property of these two IQAs. Since both algorithms regard the quality score (QS) 100 as the worst quality, QS 0 as the best quality, we convert the QS using QS_{new}=1-QS/100 and compare QS_{new} from the two algorithms with the predictions of rFDA from our proposed algorithm. The correlation results are presented in Table 7. For the proposed model, compared with the results obtained from the reduced testing set (shown in Table 6), the correlation coefficients on this complete testing set are slightly higher. This indicates that the proposed model can also achieve good performance on the original videos and the down-

Figure 8. Full-reference IQAs performance



(a) rFDA VS. PSNR (b) rFDA VS. SSIM

Table 6. Full-reference correlation coefficients

Algorithms	LCC	KRCC	SROCC
SSIM	0.4711	0.3049	0.4619
PSNR	0.4905	0.3191	0.4824
Proposed	**0.9130**	**0.7261**	**0.9050**

sampled versions. The correlation coefficients of BRISQUE and BLIINDS-II are quite low, among them, the maximum value is 0.1020 and the minimum value (-0.0105) is even negative. The scatter figures between BRISQUE, BLIINDS-II and rFDA values are shown in Figure 9, in which a certain level of perceptual quality indicated by BRISQUE or BLIINDS-II could correspond to a diverse range of rFDA values. These results indicate that the proposed model is a good image quality estimator of object detection for various kinds of distortion; however, BRISQUE and BLIINDS-II are limited in predicting image quality for object detection since they are intended for predicting perceptual quality.

Finally, in order to validate the generalization capability of the proposed model, 5 different combinations of 2 videos in the testing set are randomly selected from 10 videos with the left 8 videos as the training set, and the regression metrics and no-reference correlation coefficients for these 5 random selections of testing sets are shown in Table 8. From Table 8, we can find that the averages of the regression metrics and the correlation coefficients are all close to the results of the previous experiments shown in Table 5 and Table 7, and the standard deviation of the performances of the 5 random selections is also small. These results indicate that the proposed model has the generalization capability for different videos.

CONCLUSION

In this paper, we have proposed a no-reference image quality model based on a wide range of object detection algorithms that can be executed on embedded cameras. The proposed model could predict image quality for object detection by considering different types of quality degradation in the imaging process, including reduced resolution, noise, and blur. The proposed model is built based on a diverse range of scene characteristics. Utilizing easily extracted local and global features, the model achieves

Figure 9. No-reference IQAs performance

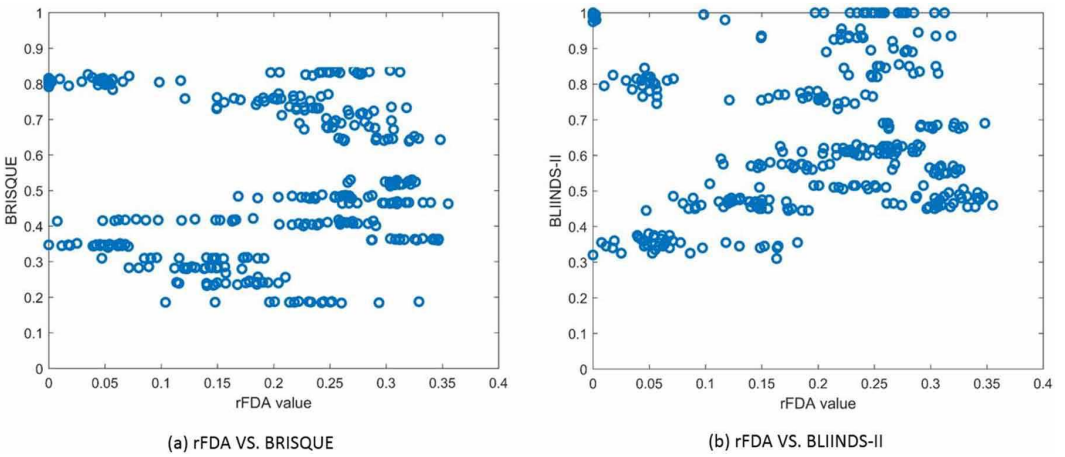


Table 7. No-reference correlation coefficients

Algorithms	LCC	KRCC	SROCC
BRISQUE	-0.0105	0.0140	0.0701
BLIINDS-II	0.0498	0.0781	0.1020
Proposed	**0.9161**	**0.7278**	**0.9094**

Table 8. Regression metrics and correlation coefficients for 5 random selections of testing set

Metrics	RMSE	R^2	$adjR^2$	MSE	MAE	LCC	KRCC	SROCC
Average	0.0455	0.7915	0.7894	0.0021	0.0358	0.8946	0.6933	0.8831
Standard deviation	0.0023	0.0190	0.0212	0.0002	0.0025	0.0132	0.0204	0.0136

more accurate predictions of image quality for object detection than common full-reference image quality measures, such as PSNR and SSIM, and popular no-reference IQAs.

In the future, we will propose a quality adjustment framework to optimize the quality of images for object detection during the image sensing process. Based on the proposed quality model, the framework will predict the performance of object detection on a sensed image. If the quality of the image is not satisfactory, pre-processing methods for removing noise or blur will be applied to enhance its quality. The proposed framework will benefit a wide range of imaging applications that rely on automatic analysis components.

REFERENCES

Abaza, A., Harrison, M. A., & Bourlai, T. (2012, November). Quality metrics for practical face recognition. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)* (pp. 3103-3107). IEEE.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886-893). IEEE Computer Society. doi:10.1109/CVPR.2005.177

Ding, C., & Tao, D. (2018). Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 1002–1014. doi:10.1109/TPAMI.2017.2700390 PMID:28475048

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 743–761. doi:10.1109/TPAMI.2011.155 PMID:21808091

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*(2), 303–338. doi:10.1007/s11263-009-0275-4

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1627–1645. doi:10.1109/TPAMI.2009.167 PMID:20634557

Gunasekar, S., Ghosh, J., & Bovik, A. C. (2014). Face detection on distorted images augmented by perceptual quality-aware features. *IEEE Transactions on Information Forensics and Security*, *9*(12), 2119–2131. doi:10.1109/TIFS.2014.2360579

Irvine, J. M., & Nelson, E. (2009, May). Image quality and performance modeling for automated target detection. In *Automatic Target Recognition XIX* (p. 73350L). International Society for Optics and Photonics. doi:10.1117/12.818593

Irvine, J. M., & Wood, R. J. (2013, May). Real-time video image quality estimation supports enhanced tracker performance. In Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications X (Vol. 8713, p. 87130Z). International Society for Optics and Photonics. doi:10.1117/12.2016174

Irvine, J. M., Wood, R. J., Reed, D., & Lepanto, J. (2013, October). Video image quality analysis for enhancing tracker performance. In 2013 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) (pp. 1-9). IEEE. doi:10.1109/AIPR.2013.6749326

Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., & Zhang, J. et al. (2009). Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 319–336. doi:10.1109/TPAMI.2008.57 PMID:19110496

Kong, L., Dai, R., & Zhang, Y. (2016, September). A new quality model for object detection using compressed videos. In *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)* (pp. 3797-3801). IEEE. doi:10.1109/ICIP.2016.7533070

Kong, L., Ikusan, A., Dai, R., & Zhu, J. (2019, March). Blind Image Quality Prediction for Object Detection. In *Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 216-221). IEEE. doi:10.1109/MIPR.2019.00046

Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., & Zhang, L. (2016). Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, *26*(2), 1004–1016. doi:10.1109/TIP.2016.2631888 PMID:27893392

Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., & Zuo, W. (2018). End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, *27*(3), 1202–1213. doi:10.1109/TIP.2017.2774045 PMID:29220321

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). MOT16: A benchmark for multi-object tracking.

Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, *21*(12), 4695–4708. doi:10.1109/TIP.2012.2214050 PMID:22910118

Nam, W., Dollár, P., & Han, J. H. (2014). Local decorrelation for improved pedestrian detection. In Proceedings of the Advances in Neural Information Processing Systems (pp. 424-432).

Pulecio, C. G. R., Benítez-Restrepo, H. D., & Bovik, A. C. (2017, September). Image quality assessment to enhance infrared face recognition. In *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)* (pp. 805-809). IEEE. doi:10.1109/ICIP.2017.8296392

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016, October). Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision* (pp. 17-35). Cham: Springer. doi:10.1007/978-3-319-48881-3_2

Saad, M. A., Bovik, A. C., & Charrier, C. (2012). Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, *21*(8), 3339–3352. doi:10.1109/TIP.2012.2191563 PMID:22453635

Tahboub, K., Reibman, A. R., & Delp, E. J. (2017, September). Accuracy prediction for pedestrian detection. In *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)* (pp. 4192-4196). IEEE. doi:10.1109/ICIP.2017.8297072

Wang, S., Gu, K., Zhang, X., Lin, W., Ma, S., & Gao, W. (2018). Reduced-reference quality assessment of screen content images. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(1), 1–14. doi:10.1109/TCSVT.2016.2602764

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. doi:10.1109/TIP.2003.819861 PMID:15376593

Yang, K. F., Li, H., Li, C. Y., & Li, Y. J. (2016). A unified framework for salient structure detection by contour-guided visual search. *IEEE Transactions on Image Processing*, *25*(8), 3475–3488. doi:10.1109/TIP.2016.2572600 PMID:27244740

Zhong, C., & Reibman, A. R. (2018). Prediction system for activity recognition with compressed video. *Electronic Imaging,* (2), pp. 1-6.

ENDNOTES

[1] This work was supported by the National Institute of Standards and Technology under Grant 60NANB17D193 and the National Science Foundation under Grant CNS-1644946.

Lingchao Kong received a M.S. degree in control engineering from Harbin Institute of Technology, Harbin, China, in 2013. He is currently working toward his Ph.D. degree in the Department of Electrical Engineering and Computer Science at University of Cincinnati. His research interests include video processing, video coding, and multimedia communications.

Ademola Ikusan is a PhD student in the Department of Electrical Engineering and Computer Science at University of Cincinnati. He received his MSc in Cybersecurity from Wright State University in 2017 and BSc in Computer Engineering from Covenant University, Nigeria in 2012. His current research is in video/image processing and quality enhancement in multimedia, network security, and networking application in multimedia.

Rui (April) Dai is an Assistant Professor in the Department of Electrical Engineering and Computer Science at University of Cincinnati, Ohio, USA. She received her BS in Electronics and Information Engineering and MS in Communications and Information Systems from Huazhong University of Science and Technology, Wuhan, China, in 2004 and 2007, respectively. She received her PhD degree in Electrical and Computer Engineering from Georgia Institute of Technology, Atlanta, GA, USA in 2011, under the supervision of Professor Ian F. Akyildiz in the Broadband Wireless Networking Laboratory. She was a postdoctoral fellow at the Center for Assistive Technology and Environmental Access of Georgia Tech in 2012. She was an Assistant Professor at the Department of Computer Science at North Dakota State University from 2012 to 2014. Her recent research interests include multimedia communications and networking, wireless sensor networks, and cyber-physical systems.

Jingyi Zhu is an undergrad student majored in Electrical Engineering in Chongqing University and University of Cincinnati Joint Dual degree Program. She will receive her B.S in 2020. She was an exchange student in Sciences Po, France and University of Queensland, Australia in Fall 2016 and Spring 2019, respectively. Her research interests are video quality enhancement, medical image processing, and machine learning.

Dara Ros is currently working toward his MS degree at the University of Cincinnati, OH, USA. He received a bachelor's degree in electrical engineering from the University of Cincinnati in 2016. His research interests include multimedia communications, digital communication, and image processing.