2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)

Blind Image Quality Prediction for Object Detection

Lingchao Kong, Ademola Ikusan, Rui Dai, and Jingyi Zhu
Department of Electrical Engineering and Computer Science
University of Cincinnati, Cincinnati, OH 45221, USA
konglo@mail.uc.edu, ikusanaa@mail.uc.edu, rui.dai@uc.edu, zhujy@mail.uc.edu *

Abstract

Automatic video data analysis tools have become indispensable components in today's imaging applications. The accuracy of automatic analysis methods relies on the quality of images or videos that are processed. It is therefore essential to introduce objective metrics for predicting the quality of images as evaluated by automatic analysis algorithms. Object detection is the first and the most important step in the process of automatic video analysis. This paper proposes a new image quality model for predicting the performance of object detection. A video data set is constructed that considers different factors related to quality degradation in the imaging process, such as reduced image 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 such as PSNR and SSIM.

1. Introduction

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 todays 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 [12, 18, 19].

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. For example, the problem of evaluating motion imagery quality for tracking in airborne reconnaissance systems was studied in [8]. 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 [11], 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 [17, 21], 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

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

978-1-7281-1198-8/19/$31.00 ©2019 IEEE
DOI 10.1109/MIPR.2019.00046

(a) MOT15-02 (b) MOT17-02 (c) MOT17-04 (d) MOT17-10 (e) MOT17-13

(f) DMcam01 (g) DMcam02 (h) DMcam04 (i) DMcam06 (j) DMcam08

Figure 1. Snapshots of video data set.

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 on a separate test data set and compared against commonly used IQA measures.

2. 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 [1, 6, 15]. 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 [1]. In [15], natural scene statistics was used to detect degradation of infrared images for face recognition. In [6], 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 [7], 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 [8] and [9], 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.

3. 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 [13], and 5 videos are chosen from the Duke Multi-Target Multi-Camera Tracking (DM) dataset [16]. The resolutions of these videos are mostly 1920×1080 except for one video with 640×480 resolution, and the average number of frames is 741. The snapshot of these videos are shown in Fig. 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. 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. For each original video sequence shown in Fig. 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) [3];

(2) Discriminatively Part Models (DPM) [5];

(3) Locally Decorrelated Channel Features (LDCF) [14].

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 [10]. 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

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

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 [4]. 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,

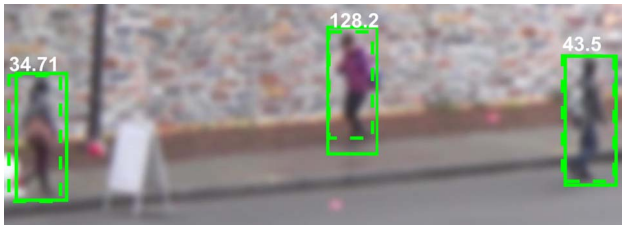


Figure 2. Detection sample.

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

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 is $T(j)$, and $T(j) \in \{C_1, ..., C_j, ..., C_{N_m}\}$.

The original FDA measure in (1) 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 Fig. 2, 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.

We have visualized the distribution of rFDA values for the images in our data set. Fig. 3 compares the rFDAs obtained from the original images and the images with high and low distortions. Fig. 3 (a), (b), (c), and (d) correspond to the results on down-sampled, blurred, motion blurred, and noisy images, 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.

4. Blind model for prediction rFDA

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

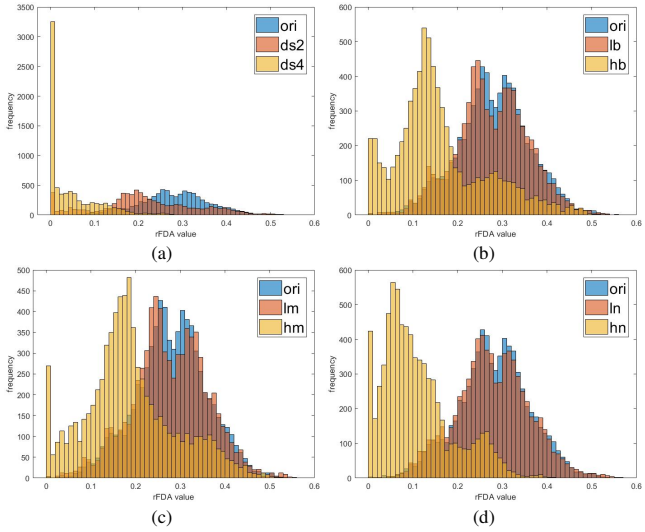


Figure 3. Distribution of rFDA on images with different types of distortions.

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 $f(x, y)$, the gradient of f at location (x, y) is defined as the two dimensional column vector: $[\partial f/\partial x, \partial f/\partial y]^T$, where $\partial f/\partial x = f(x+1, y) - f(x-1, y)$, and $\partial f/\partial y = f(x, y+1) - f(x, y-1)$ using finite difference filters. The magnitude and direction of this gradient at location (x, y) are given by

$$mag(x) = \sqrt{(\partial f/\partial x)^2 + (\partial f/\partial y)^2}, \tag{3}$$

$$dir(x, y) = tan^{-1}\left[\frac{\partial f/\partial y}{\partial f/\partial x}\right]. \tag{4}$$

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.

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 [3]. The local window for one HOG descriptor is set as 16×16 pixels, and 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:

$$w_m = \sum_{i=1}^{N_b} h_i/N_b, \tag{5}$$

$$w_s = \sqrt{\sum_{i=1}^{N_b} (h_i - w_m)^2/(N_b - 1)}, \tag{6}$$

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 3x3 kernels in the horizontal and vertical directions. The local 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 in [20]. First, a contour-based spatial prior is extracted based on the layout of edges in the given image along a non-selective pathway. Then, local features such as color, luminance, and texture, are extracted along the selective pathway. Finally, Bayesian inference is used to auto-weight and integrate the local cues to predict the exact locations of objects.

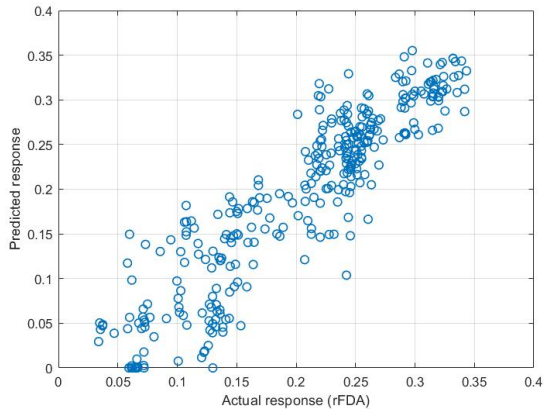
We use the bootstrap aggregating, or bagging, ensemble of trees to train a regression model to predict detection performance based on the extracted 13 features on a single image [2]. Every decision tree in the bagging ensemble is grown on an independently drawn bootstrap replica of input observations. 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.

5. Performance evaluation

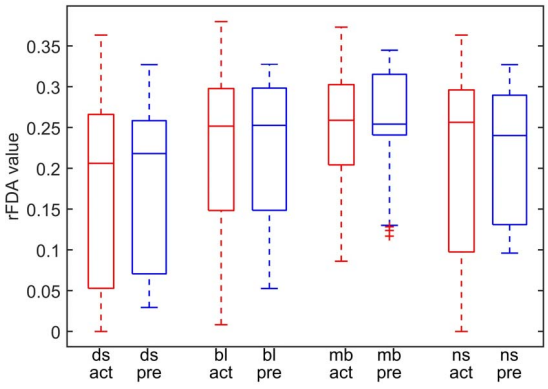
Category	video name	image number	percentage
Training set	MOT17-02, MOT17-10, MOT15-02, DMcam01, DMcam02, MOT17-13, DMcam04, DMcam08	50022	75.03%
Testing set	MOT17-04, DMcam06	16650	24.97%

Table 1. Learning setting

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 1. The total number of images



(a) Actual response VS. predicted response



(b) Distribution comparison in distortion category

Figure 4. Regression performance.

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. During training, 5-fold cross validation, 30 base learners, and a minimum leaf size of 8 are used to build the ensemble of trees.

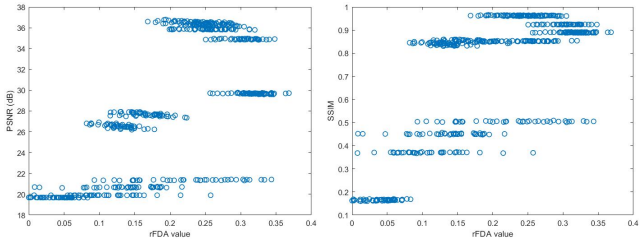
First, the regression performance of the proposed model is investigated. Fig. 4 (a) shows the scatter figure of the actual response VS. the predicted response. There are a huge number of observations (16650 images) in the testing data set, and one point is selected from every 50 observations to generate 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. Fig. 4 (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.

Metrics	RMSE	R^2	$adjR^2$	MSE	MAE
Values	0.0428	0.8147	0.8146	0.0018	0.0324

Table 2. Regression metrics

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 2. 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 2, we can find that the values of RMSE, MSE and MAE are 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 performance of the proposed model is also compared with popular 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 3. 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 can not be good indicators for the image quality for object detection. The conclusion also can be drawn from Fig. 5, in which scatter figures between PSNR, SSIM and rFDA values are plotted. From Fig. 5, 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.



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

Figure 5. IQAs performance.

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

Table 3. Correlation coefficients

6. Conclusion

In this paper, we have proposed a blind image quality model for a wide range of object detection algorithms and diverse scene characteristics. Utilizing easily extracted local and global features, the model achieves more accurate predictions of image quality for object detection than common image quality measures such as PSNR and SSIM.

References

[1] A. Abaza, M. A. Harrison, and T. Bourlai. Quality metrics for practical face recognition. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3103–3107. IEEE, 2012.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[4] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[6] S. Gunasekar, J. Ghosh, and A. C. Bovik. Face detection on distorted images augmented by perceptual quality-aware features. *IEEE transactions on information forensics and security*, 9(12):2119–2131, 2014.

[7] J. M. Irvine and E. Nelson. Image quality and performance modeling for automated target detection. In *Automatic Target Recognition XIX*, volume 7335, page 73350L. International Society for Optics and Photonics, 2009.

[8] J. M. Irvine and R. J. Wood. Real-time video image quality estimation supports enhanced tracker performance. In *Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications X*, volume 8713, page 87130Z. International Society for Optics and Photonics, 2013.

[9] J. M. Irvine, R. J. Wood, D. Reed, and J. Lepanto. Video image quality analysis for enhancing tracker performance. In *Applied Imagery Pattern Recognition Workshop (AIPR): Sensing for Control and Augmentation, 2013 IEEE*, pages 1–9. IEEE, 2013.

[10] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. 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, 2009.

[11] L. Kong, R. Dai, and Y. Zhang. A new quality model for object detection using compressed videos. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 3797–3801. IEEE, 2016.

[12] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2018.

[13] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[14] W. Nam, P. Dollár, and J. H. Han. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2014.

[15] C. G. R. Pulecio, H. D. Benítez-Restrepo, and A. C. Bovik. Image quality assessment to enhance infrared face recognition. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 805–809. IEEE, 2017.

[16] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[17] K. Tahboub, A. R. Reibman, and E. J. Delp. Accuracy prediction for pedestrian detection. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 4192–4196. IEEE, 2017.

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

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[20] K.-F. Yang, H. Li, C.-Y. Li, and Y.-J. Li. A unified framework for salient structure detection by contour-guided visual search. *IEEE Transactions on Image Processing*, 25(8):3475–3488, 2016.

[21] C. Zhong and A. R. Reibman. Prediction system for activity recognition with compressed video. *Electronic Imaging*, 2018(2):1–6, 2018.