ACCURACY PREDICTION FOR PEDESTRIAN DETECTION

Khalid Tahboub* *Amy R. Reibman*† *Edward J. Delp**

* Video and Image Processing Lab (VIPER), Purdue University, West Lafayette, Indiana USA
† School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana USA

ABSTRACT

In this paper, we address the problem of predicting accuracy for pedestrian detection. We want to be able to predict the accuracy of a video analytic method without actually executing the method. We propose the use of texture descriptors and random forests to predict the accuracy of various pedestrian detection methods. Our experimental results demonstrate that using the local binary pattern (LBP) or a bank of Schmid and Gabor filters can capture spatial textural information associated with video quality degradation that can be used to predict accuracy. We also demonstrate how predicting the absolute accuracy can save network and computational resources.

***Index Terms*—** Pedestrian detection, accuracy prediction, video quality, video compression

1. INTRODUCTION

Video surveillance systems are widely deployed with the number of cameras increasing exponentially [1]. Due to the large number of cameras, continuous human monitoring is no longer possible in most surveillance system with the video usually archived for forensic purposes.

The use of video analytics to automatically identify activities and anomalous behaviors is very important [2, 3]. The accuracy of video analytics depends on the input video quality. In mobile or networked environment, bandwidth is limited and adaptive data-rate video streaming is used [4]. Video compression can introduce quality degradation that impacts the accuracy of video analytics. The impact of MJPEG compression on the performance of tracking vehicles was investigated in [5]. In [6], compression and frame rate were varied and the impact on video analytics were investigated. Recommendations for acceptable data-rates with applications in face recognition were proposed in [7]. These methods characterize the impact of compression on analytics but do not attempt to predict the accuracy when the original compression level is unknown.

The ability to predict the accuracy of video analytics can save network and computational resources. At the video encoder, the quantization parameter (QP) can be increased as long as the predicted performance does not deteriorate significantly. This uses network resources efficiently but requires the encoder to know which analytics method will be used. At the decoder, where video analytics reside, the ability to predict the performance of various methods allows one to select the least computationally expensive method that satisfies the system end-to-end performance requirements. We want to be able to predict the accuracy of a video analytic method without actually executing the method.

To do this we to need measure video quality. Traditional video quality assessment methods aim to measure quality as perceived by humans [8]. The quality of experience is the main drive behind such metrics and cannot be generalized to video analytics. Current quality models for video analytics are not adequate to predict the accuracy. In [9], a quality model was proposed for moving object detection. Three moving object detection methods were selected. The goal was to estimate the relative performance of the three methods on compressed video sequences in comparison to the output when no compression is used. The sum-of-absolute frame difference in macro block (MB) units was used to estimate number of false positives (FP), and absolute difference of texture in MB units was used to estimate the number of false negatives (FN). However, there was no attempt to predict the accuracy. In [10], a quality model to predict object tracking performance in video sequences was proposed. To extend traditional performance metrics to multi-target object tracking, they defined generalized target detections and generalized false alarms. Frame rate and spatial resolution were changed and the impact on performance was investigated. Entropy, autocorrelation function across frames, Laplacian of the image and noise metrics were used as image quality metrics and showed clear relationship with the tracker performance. However, tracking targets at multiple scales was not addressed and only a single tracker was investigated.

In this paper, we address the problem of accuracy prediction for one type of video analytic, pedestrian detection. We propose the use of texture descriptors and random forest to predict the accuracy. We examine our proposed model using four pedestrian detectors using compressed video sequences. Since the detection rate is also dependent on the target scale, we show that our model is able to predict the miss rate when the goal is to detect pedestrians larger than 80 pixels or pedestrians larger than 50 pixels. We also propose practical use cases in which our prediction method can save network and computational resources.

2. PEDESTRIAN DETECTION

Detecting humans is at the core of surveillance systems, action recognition tasks and autonomous driving. The performance of pedestrian detectors have improved by adopting new learning techniques, the use of new hand-crafted features and representation learning techniques [11]. Designing shape descriptors based on edges or gradients is a common approach for pedestrian detection [12, 13]. Convolutional neural network (CNN) methods have also shown significant gains [14–17]. Based on a comparison with a human baseline, failure cases are clustered in two categories: significantly occluded pedestrians and small scales [17]. Some of the publicly available datasets include: ETH dataet [18], Daimler detection benchmark (Daimler-DB) [19] and Caltech pedestrian dataset [20].

In this paper, we propose a model to predict the accuracy of pedestrian detection and test it using four detectors:

Histogram of oriented gradients (HOG) [12] uses normalized

This work was partially supported by the U.S. Department of Homeland Security's VACCINE Center under Award number 2009-ST-061-CI0001 and the Cisco University Research Program Fund CG-#594368 through the Silicon Valley Community Foundation. Address all correspondence to Edward J. Delp, ace@ecn.purdue.edu

local histograms of image gradients orientations in a dense grid. Human shape is characterized by the distribution of edge orientations [12]. A linear support vector machine (SVM) is used for classification.

Aggregated channel features (ACF) [21] uses a multi-scale representation with fast feature pyramids. Adaptive Boosting is used to train and combine decision trees with a multi-scale sliding-window.

Locally decorrelated channel features (LDCF) [22] reduces ACF miss rate by using decorrelating (linear) filters. Four filters per channel are applied to the original ten channels of ACF.

Deformable part model (DPM) [23] also uses a multi-scale representation. It is based on the idea that an object can be represented by parts. HOG coarse features are used to represent an object while finer high resolution features are used to represent parts. The spatial model follows a star graph.

We use the log-average miss rate to determine the performance of pedestrian detectors. It is computed by averaging the miss rate at nine false positive per image (FPPI) rates evenly spaced in the $(0.01 - 1)$ range [20]. Throughout the paper, we will refer to the log-average miss rate simply as the miss rate. Experiments are conducted on the Caltect pedestrian dataset [20]. It is comprised of approximately 250,000 frames in 137 minute long segments. Video spatial resolution is 640×480 at 30Hz captured from a vehicle driving through an urban environment. A total of 350,000 bounding boxes were annotated for 2300 unique pedestrians who vary in scale, level of occlusion and location. The dataset is divided into 11 sessions: $(S0-S5)$ are typically used for training and $(S6-S10)$ for testing. We refer to $(S0-S5)$ as $\text{training}_{\text{Ped}}$ and use it to train pedestrian detectors. We divide $(S6 - S10)$ into two parts: 1) $\text{training}_{\text{QM}}$ is used to train the quality model 2) $\text{testing}_{\text{Red}}$ is a reduced version of the original testing dataset and is used for experiments. We used FFmpeg [24] to transcode the whole dataset using H.264 format at 30 frames/second. The dataset was transcoded 18 times, each time using a unique quantization parameter (QP). The set of QPs is: $\{15, 20, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55\}$

To investigate the impact of video compression, we evaluate the ACF and LDCF detectors using compressed video sequences. We obtain the detection results assuming the goal is to detect pedestrians larger than 50 pixels and larger than 80 pixels. Throughout the paper, we will refer to detecting pedestrians larger than 50 pixels as P50 and detecting pedestrians larger than 80 pixels as P80. For each detector and each detection goal, the miss rate is computed for each transcoded version of Caltech $\text{training}_{\text{QM}}$ dataset. Figure 1 depicts this relationship for the ACF and LDCF detectors. As expected, miss rate increases with QP due to video quality degradation. In addition, when considering a particular detector miss rate values are always smaller when P80 is the detection goal. LDCF performance is also consistently better.

We also observe that the detection goal of P50 is more sensitive to video quality degradation. For example, increasing the QP value from 15 to 43 results in more than 10% increase in the miss rate for P50, whereas for P80, the QP value can be increased up to 47 to introduce 10% increase in the miss rate. Therefore, to predict the accuracy of pedestrian detection, video quality needs to be measured and the detection goal has to be stated.

3. PROPOSED MODEL

Our accuracy prediction has three components: texture descriptor extraction, spatial and temporal averaging and random forest regression. The block diagram is shown in Figure 2. Our goal is to capture large-scale spatial textural information associated with video quality degradation that can be used to predict accuracy. We explore the use of various texture descriptors as predictors. Averaging the descriptors spatially and temporally is necessary to capturing the texture associated with quality degradation as opposed to scene specific content. The random forest regression model estimates the relationship between texture descriptors and the accuracy of pedestrian detection. It is trained to predict the absolute accuracy for each one of the four pedestrian detectors and for both detection goals: P50 and P80.

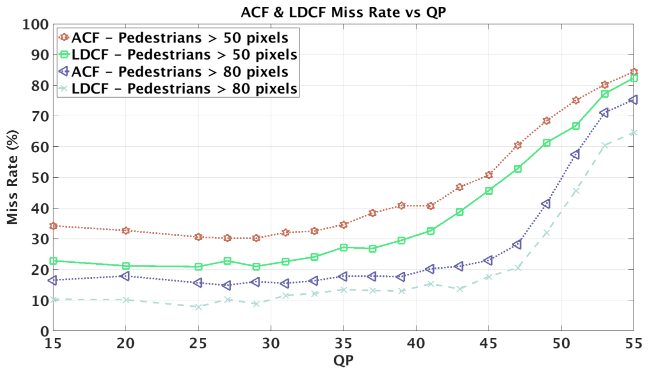


Fig. 1. ACF and LDCF miss rate versus QP

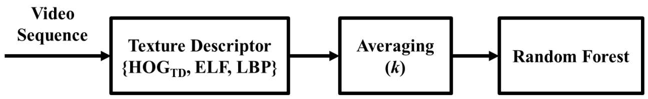


Fig. 2. Block diagram of our approach

3.1. Texture Descriptors

The original Caltech images are 640×480 pixels, texture descriptors are extracted from 128×128 image windows. In this paper, we explore the use of three texture descriptors:

Histogram of oriented gradients (HOG) [12], as mentioned before, uses normalized local histograms of image gradients orientations in a dense grid. Cell size, block size and the number of histogram bins are the main parameters associated with the HOG descriptor. Since our goal is to capture large-scale spatial information associated with video quality degradation, we use 16×16, 32×32 and 64×64 cell sizes. For each cell size, we extract the HOG descriptor using 9 bins per histogram and with a block comprised of one cell. We concatenate the three descriptors to construct a 756-D vector. We will refer to this vector as HOG_{TD} to avoid any confusion with the pedestrian detector.

Local binary pattern (LBP) [25] determines a binary code for each pixel in an image and uses the histogram of binary codes as a texture descriptor. To compute the binary code, the pixel value is used as a threshold and compared with each neighbor value. The neighbors are selected from a circular pattern centered at the pixel. The radius of the circle, number of neighbors and the cell size over which the histogram is accumulated are the main parameters associated with this process. We use six descriptors from 16 or 32 neighbors in combination with 16, 32 or 63 pixels as radius. The six descriptors are concatenated to construct a 156-D vector.

Ensemble of localized features (ELF) was proposed to model texture for pedestrian re-identification [26]. We adopt the same approach to predict the accuracy of pedestrian detectors. 13 Schmid filters [27] are used to model rotation invariant texture and 8 Gabor filters [28] are used with vertical and horizontal strips. Sixteen bin histograms are constructed for each of the 21 filter responses. The histograms are concatenated to form a high dimensional vector with 336 features.

3.2. Random Forest

Each of the texture descriptors mentioned above is a high dimensional feature vector. The features within the texture descriptors vary in terms of their importance. Some of the features are irrelevant and some are important to capturing information associated with video quality degradation. We propose the use of random forest due to its automatic ability to select features, overcome overfitting and discard irrelevant features [29]. Random forest is based on a bootstrap-aggregated ensemble of decision trees. Bootstrap aggregation refers to generating multiple training datasets out of the original training data and using each set to train a separate tree. Training data consists of pairs of predictors and responses. We denote the original training data by D: (X_i,Y_i) where $i = 1, \ldots, n$. X's are predictors and Y's are responses. Sampling from a uniform distribution with replacement is used to generate the bootstrap samples D_l's, $l = 1, \ldots, m$. The random forest is defined as the set of trees $\{h(X, D_l), l = 1, \ldots, m\}$ and the final prediction is simply computed by averaging the m individual predictions. When growing each tree, random forest also employs randomness for selecting a variable to split on [29].

To capture spatial textural information associated with video quality degradation, we propose to use the average of multiple descriptors as a predictor. We used all transcoded versions of the Caltech training$_{\text{QM}}$ dataset. Each transcoded version contains 53,143 images. We used 1 frame out of each 10 frames (3 frames/second). 5 windows (128 × 128) were selected randomly from each frame, and a texture descriptor (T) was extracted for each one of the frame windows. A Total of 26,571 descriptors were generated for each transcoded dataset. For each k texture descriptors, one predictor was obtained:

$$X_i = \sum_{w=(i-1)\times(k)+1}^{i\times k} T_w^{\text{QP}}$$

where

T_w^{QP} is the wth descriptor for the dataset transcoded at QP,

$$i = 1, \ldots, \lfloor \tfrac{26{,}571}{k} \rfloor$$

Responses correspond to the miss rate of pedestrian detection based on the ground truth. Generating predictor/response pairs are repeated for each transcoded version of the dataset. A random forest was built for each pedestrian detector and for each detection goal: P80 and P50.

4. EXPERIMENTAL RESULTS

In this section, we conduct several experiments to evaluate our proposed method. First, we measure the accuracy by comparing the predicted miss rates to the true values. The second experiment examines the usability of our approach and its ability to adjust the QP value based on pedestrian detection performance. The third experiment investigates the parameter k and investigates how to capture textural content associated with video quality degradation. The last experiment examines if the proposed approach is able to function in real-time.

The first experiment is conducted using each transcoded version of the Caltech testing$_{\text{Red}}$ dataset. True miss rates are computed for the 4 pedestrian detectors and for both detection goals: P50 and P80. Miss rate values are also predicted using our proposed approach. Each random forest is trained using the Caltech training$_{\text{QM}}$ dataset with 8 trees. k is set to 24 and the final miss rate prediction is obtained by averaging the random forest responses over the dataset. Results are shown in Figure 3. Figure 3(a) and 3(b) show the P80 miss rate values for LDCF and ACF, respectively, whereas Figure 3(c) and 3(d) show the P50 miss rate values for HOG and DPM, respectively. Each marked point is a true or a predicted miss rate at a particular quantization parameter. The LBP and ELF texture descriptors lead to more accurate predictions as compared to the HOG$_{\text{TD}}$ descriptor. To measure how close the predictions are to the true values, we use the mean absolute error (MAE) [30]:

$$\text{MAE} = \tfrac{1}{18} \sum_{\text{QP}} |\text{TMR} - \text{PMR}|,$$

where TMR is the true miss rate and PMR is the predicted miss rate. Table 1 summarizes the MAE's for each texture descriptor and each case. The LBP based predictions has the lowest MAE. It also has a consistent performance over both detection goals, whereas ELF based approach is more accurate in predicting miss rates for P80.

Table 1. Mean absolute errors for miss rate prediction

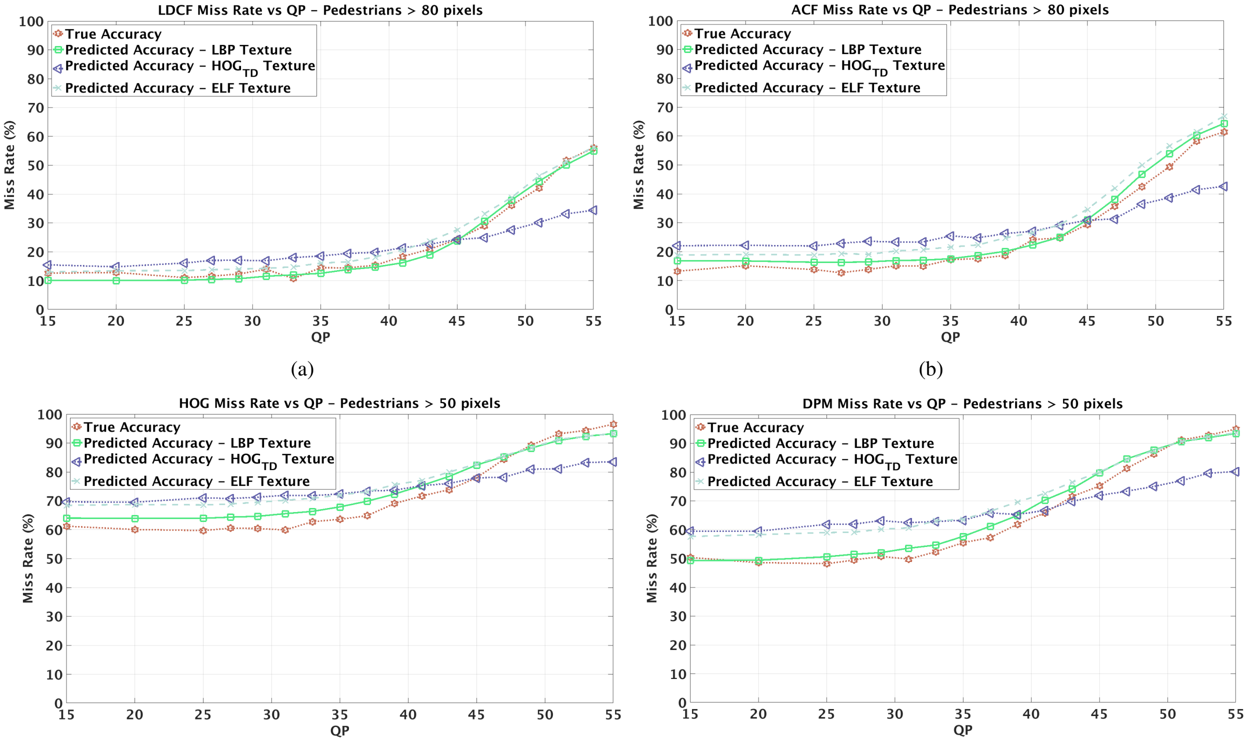
	LBP	HOG$_{\text{TD}}$	ELF
HOG - pedestrians > 80	4.4%	6.6 %	2.1%
ACF - pedestrians > 80	2.3%	8.3 %	5.2%
LDCF - pedestrians > 80	1.6%	6.3 %	2.1%
DPM - pedestrians > 80	3.9%	9.2 %	2.1%
Average - pedestrians > 80	3.0%	7.6 %	2.9%
HOG - pedestrians > 50	3.5%	8.3 %	6.1%
ACF - pedestrians > 50	4.9%	11.6%	8.5%
LDCF - pedestrians > 50	1.8%	9.7 %	5.5%
DPM - pedestrians > 50	2.3%	9.4 %	6.5%
Average - pedestrians > 50	3.1%	9.7 %	6.6%
Total average	3.1%	8.7 %	4.8%

Figure 3 also demonstrates that for LBP based predictions the pattern of predicted values follows the pattern of true values very closely. This is a significant result from a practical perspective. In a networked environment with no prior knowledge of video quality, an encoder or a transcoder device can incrementally increase the QP value as long as the predicted miss rate does not deteriorate significantly. An experiment is conducted to evaluate the accuracy of LBP based predictions in determining the QP value for which the miss rate is not increased by more than T%. To measure how close the predictions are to the true values, we use the mean absolute percentage error (MAPE) [30]:

$$\text{MAPE} = \tfrac{1}{8} \sum_{\{\text{Tests}\}} |\tfrac{\text{TQP}-\text{PQP}|}{\text{TQP}},$$

where TQP (or *PQP*) is the QP value for which the miss rate is not increased by more than T% according to the true (or *predicted*) miss rates. $\{$Tests$\}$ is the set of 8 tests conducted using 4 pedestrian detectors and 2 detection goals. MAPE is 96.7% and 96.9% when T is set to 5% and 10%, respectively. This demonstrates that our proposed model can be used in surveillance systems to efficiently control the QP value. Using this result, intelligent surveillance systems can adjust the data rate according to the detection goal, thus saving network resources. For example, if the detection goal changes from detecting pedestrians P50 to P80, the data rate can be lowered while maintaining the same predicted miss rate.

The third experiment investigates the impact of changing the number of texture descriptors (k) used to compute a single predictor. k is increased from 1 to 50 in increments of 4. For each k value, the MAE is averaged over all use cases and all pedestrian detectors. In Figure 4, we plot the MAE as a function of k for predictions based on LBP and ELF texture descriptors. The optimal range for k is between 5 and 45. By choosing a k value in this range, we are able



(a)

(b)

(c)

(d)

Fig. 3. Prediction results

to capture textural information associated with video quality degradation as opposed to very low k values where it is associated with scene specific content.

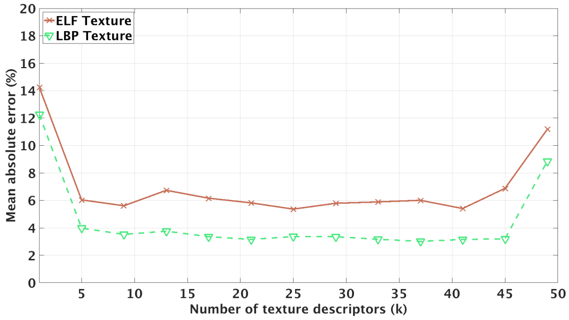


Fig. 4. Mean absolute error as a function of k.

The practical applications mentioned above require the ability to make predictions in real-time. In the last experiment, we investigate the impact of using a finite number of random forest responses to predict the miss rate of the whole dataset. MAE was computed and averaged over the 4 pedestrian detectors and both detection goals. In Figure 5, MAE is plotted as a function of the number of responses, error bars represent one standard deviation. As expected, the error rate decreases rapidly with the number of responses. Since the texture descriptors are extracted from 128 × 128 image windows, it is plausible to compute a large number of responses from the same frame making our proposed method suitable for on-line decisions.

5. CONCLUSIONS

In this paper, we proposed a method to predict the accuracy of various pedestrian detection methods without needing to actaully execute the methods. Our experimental results demonstrated that the use of the local binary pattern (LBP) achieves the best accuracy in predicting the miss rate for multiple pedestrian detectors and for two different detection goals. The results also demonstrated that averaging multiple descriptors are necessary to capture spatial textural information associated with video quality degradation as opposed to scene-specific content. We also demonstrated practical use cases in which our prediction method can save network and computational resources.

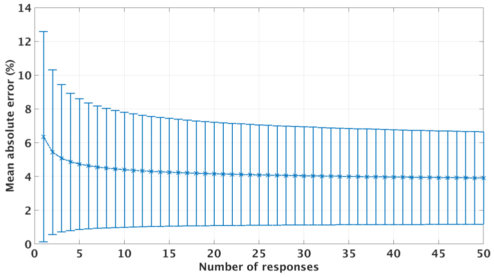


Fig. 5. Mean absolute error for various number of responses.

6. REFERENCES

[1] N. Haering, P. L. Venetianer, and A. Lipton, "The evolution of video surveillance: an overview," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 279–290, October 2008.

[2] W. Li, "Anomaly detection and localization in crowded scenes," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1975–1981, June 2010, San Francisco, California.

[3] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Journal of Computing Surveys*, vol. 43, no. 3, pp. 1–43, April 2011.

[4] K. Tahboub and E. J. Delp, "Chicago LTE video pilot final lessons learned and test report," October 2015, Available at: https://www.dhs.gov/publication/chicago-lte-video-pilot-report'.

[5] A. Cozzolino, F. Flammini, V. Galli, M. Lamberti, G. Poggi, and C. Pragliola, "Evaluating the effects of mjpeg compression on motion tracking in metro railway surveillance," *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 142–154, September 2012, Brno, Czech Republic.

[6] A. Tsifouti, S. Triantaphillidou, M. C. Larabi, G. Doré, and A. Psarrou, "The effects of scene content parameters, compression, and frame rate on the performance of analytics systems," *Proceedings of the SPIE Conference on Image Quality and System Performance*, p. 93960X, February 2015, San Francisco, CA.

[7] A. Tsifouti, S. Triantaphillidou, E. Bilissi, and M. C. Larabi, "Acceptable bit-rates for human face identification from CCTV imagery," *Proceedings of the SPIE Conference on Image Quality and System Performance*, p. 865305, February 2013, Burlingame, CA.

[8] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf, "Objective video quality assessment system based on human perception," *Proceedings of the IS&T/SPIE Conference on Human Vision, Visual Processing, and Digital Display*, pp. 15–26, January 1993, San Jose, CA.

[9] L. Kong, R. Dai, and Y. Zhang, "A new quality model for object detection using compressed videos," *Proceedings of the IEEE International Conference on Image Processing*, pp. 3797–3801, September 2016, Phoenix, AZ.

[10] J. M. Irvine and R. J. Wood, "Real-time video image quality estimation supports enhanced tracker performance," *Proceedings of the SPIE Conference on Defense, Security, and Sensing*, p. 87130Z, May 2013, Baltimore, MD.

[11] K. Yang, E. J. Delp, and E. Du, "Categorization-based two-stage pedestrian detection system for naturalistic driving data," *Signal, Image and Video Processing*, vol. 18, no. 1, pp. 135–144, October 2014.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886–893, June 2005, San Diego, CA.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[14] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1904–1912, December 2015, Santiago, Chile.

[15] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" *Proceedings of the IEEE European Conference on Computer Vision*, pp. 443–457, October 2016, Amsterdam, Netherlands.

[16] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5079–5087, June 2015, Boston, Massachusetts.

[17] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1259–1267, June 2016, Las Vegas, NV.

[18] A. Ess, B. Leibe, and L. V. Gool, "Depth and appearance for mobile scene analysis," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–8, October 2007, Rio de Janeiro, Brazil.

[19] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, October 2009.

[20] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, August 2012.

[21] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, August 2014.

[22] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," *Proceedings of the Advances in Neural Information Processing Systems Conference*, pp. 424–432, December 2014, Montréal, Canada.

[23] 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*, vol. 32, no. 9, pp. 1627–1645, September 2010.

[24] "FFmpeg," URL: http://www.ffmpeg.org.

[25] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, July 2002.

[26] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *Proceedings of the 10th European Conference on Computer Vision*, pp. 262–275, October 2008, Marseille, France.

[27] C. Schmid, "Constructing models for content-based image retrieval," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 39–45, December 2001, Kauai, HI.

[28] I. Fogel and D. Sagi, "Gabor filters as texture discriminator," *Biological cybernetics*, vol. 61, no. 2, pp. 103–113, June 1989.

[29] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, October 2001.

[30] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, October 2006.