QUALITY-ADAPTIVE DEEP LEARNING FOR PEDESTRIAN DETECTION

Khalid Tahboub[⋆] *David Güera*[⋆] *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

Pedestrian detection is a fundamental task for many applications including autonomous vehicles and surveillance systems. In a mobile or networked environment bandwidth is limited and adaptive data-rate streaming is used. Video compression can introduce significant quality degradation that impacts the accuracy of video analytics. In this paper, we examine the problem of a changing video data-rate and examine how it affects the performance of video analytics, in particular pedestrian detection, using a two-stage quality-adaptive convolutional neural network system. Our experimental results demonstrate that when adaptive data-rate streaming is used, our proposed quality-adaptive approach reduces the miss rate by 20% compared to the baseline detector.

***Index Terms*—** Pedestrian detection, deep learning, adaptive data-rate streaming, intelligent video surveillance

1. INTRODUCTION

In recent years, the number of video surveillance systems has increased dramatically. A study released by Cisco predicts that Internet video surveillance traffic will increase tenfold between 2015 and 2020. It has been predicted that 3.9% of all Internet video traffic will be due to video surveillance in 2020 an increase from 1.5% in 2015 [1].

Video analytics will require a networked infrastructure [2] using adaptive data-rate streaming [3]. Video compression can introduce significant quality degradation which impacts the performance of video analytics.

Surveillance systems, autonomous driving and action recognition are examples where human detection (e.g. pedestrian detection) plays an essential role [4, 5]. The use of convolutional neural network (CNN) methods has resulted in significant gains in human detection [6–8]. Detecting human parts [9], the use of region proposal networks [10] and incorporating semantic attributes [11] have shown great success. Despite recent progress in pedestrian detection, there are still issues associated with video quality degradation.

Some existing methods have characterized the impact of compression on analytics. The impact of MJPEG compression on tracking performance is investigated in [12]. It is shown that the performance of tracking moving objects degrades as the data rate decreases. Recommendations for acceptable data-rates for applications in face recognition are proposed in [13]. In [14], compression and frame rate are varied and the impact on the performance of video analytics is investigated. In [15, 16] we investigate the impact of video quality degradation on a method used for crowd flow estimation. An alternative approach is to extract image features using camera nodes and to send a compressed feature descriptor over the network. This is known as analyze-then-compress (ATC). In [17], ATC paradigm is compared against the traditional approach, compress-then-analyze (CTA), where images are compressed and sent over the network for analytics. The ATC paradigm obtains better results at low data-rates, while at high data-rates the CTA paradigm is the preferred choice.

In this paper, we propose a two-stage quality-adaptive convolutional neural network to address the problem of variable video data-rate. We characterize the impact of a variable video data-rate using a popular pedestrian detector [10] and investigate the use of compressed images for training. The first stage of our proposed method estimates video quality while the second stage performs pedestrian detection. Note that the first stage is not estimating quality from the perspective of human perception [18], but quality for analytics. The second stage consists of a bank of neural networks each of which is trained using images from a video sequence compressed using advanced video coding (AVC/H.264) at a unique quantization parameter. We assume that the encoder adapts to the changing bandwidth by changing the quantization parameter.

2. PEDESTRIAN DETECTION USING COMPRESSED VIDEO SEQUENCES

Many methods have been proposed and have utilized various features and learning techniques in video analytics. The use of gradient-based features results in significant performance gains as demonstrated by the histogram of oriented gradients (HOG) descriptor and scale-invariant feature transform (SIFT) [19–22]. Recently, Convolutional Neural Network (CNN) methods have been investigated for pedestrian detection and have achieved high performance [9, 11]. In [23], failure cases over multiple datasets are investigated and summarized into two categories: significant occlusions and small scales. Some of the most popular benchmarks include: INRIA person dataset [19], ETH dataet [24], TUD-Brussels pedestrian dataset [25], Daimler detection benchmark (Daimler-DB) [26] and Caltech pedestrian dataset [27].

Very deep convolutional networks [28] (known as VGG), Fast/Faster Region-based Convolutional Neural Network methods (R-CNN) [29, 30] have demonstrated excellent performance for large-scale object recognition. In [10], Fast/Faster R-CNN networks are analyzed and adopted for pedestrian detection. A Region Proposal Network followed by a Boosted Forest (RPN+BF) is proposed and results in a substantially better localization accuracy and achieves state-of-the-art performance. We use RPN in our method and train it using images from video sequences compressed at various data rates. RPN uses the VGG-16 network [28] which is pre-trained with ImageNet [31]. It combines VGG-16 with an intermediate 3×3 convolutional layer and two sibling 1×1 convolutional layers for classification and bounding box regression. A cascaded Boosted Forest classifier is trained using RPN region proposals, confidence scores and features.

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

To evaluate the performance of various approaches, we use the log-average miss rate. This summarizes the performance of pedestrian detectors and is obtained by averaging the miss rate of nine false positives per image (FPPI) evenly spaced in the interval (0.01 - 1) in log-space range [27]. We refer to the log-average miss rate simply as the miss rate.

Video compression introduces significant quality degradation which impacts video analytics. To demonstrate this, we evaluated the RPN detector using transcoded versions of the Caltech dataset. The Caltech pedestrian dataset is an extensive pedestrian datasets [27]. It is comprised of approximately 250,000 frames in 137 minute long segments. The 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 that vary in scale, level of occlusion and location.

The Caltech dataset is divided into 11 sessions: (S0-S5) are typically used for training and (S6-S10) for testing. We refer to (S0-S5) as training_PD and use it to train our pedestrian detector. We also divide (S6-S10) into two parts: training_QE which is used to train for quality estimation as will be described later, and testing_RD which is a reduced version of the original testing dataset and is used for our experiments. We use FFmpeg [32] to transcode the dataset using H.264 at 30 frames/second. The dataset is transcoded 18 times, each time using a unique quantization parameter (QP). The set of QPs is: $\{15, 20, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55\}$.

The performance of the RPN CNN is evaluated using transcoded versions of the Caltech testing_RD dataset. For each test, we find the miss rate for pedestrians > 50 pixels and pedestrians > 80 pixels. Throughout this paper, we refer to "larger than 50 pixels" as L50 and "larger than 80 pixels" as L80. In Figure 1, shows the miss rate as a function of the QP value. As expected, miss rate increases with QP due to video quality degradation. Each marked point in the curve is the miss rate at this particular quantization parameter. Miss rate values are always smaller for L80. For small QP, the miss rates are approximately 5% and 12% for L80 and L50, respectively. Figure 1 also shows detecting pedestrians L50 is more sensitive to video compression as reflected by the knee point. Video compression introduces degradation in performance near QP = 32 for pedestrians L50 compared to QP = 37 for pedestrians L80. Thus, the degradation in performance not only depends on the compression rate but also on what is being sought in the video sequence.

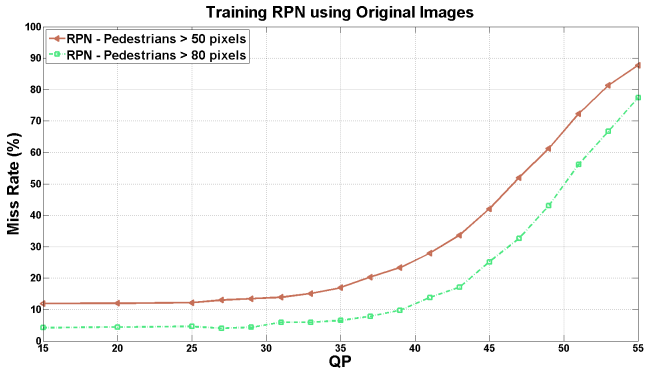

Fig. 1. Video compression and RPN performance

3. PROPOSED METHOD

The miss rate generally increases with QP. Can we achieve better performance using an RPN trained using compressed video sequences? In this section, we investigate the use of compressed video sequences to train RPN and propose a two-stage neural network to address the problem of a changing video data-rate.

3.1. Training RPN Using Compressed Video Sequences

The Caltech training_PD dataset is transcoded 18 times using 18 unique QP values. Each dataset is used to train an RPN to obtain a unique pedestrian detector. We evaluate the 18 detectors using the transcoded versions of the Caltech training_QE dataset. The miss rate values are obtained assuming the goal is to detect pedestrians L80. In Figure 2, the miss rates are shown as a function of QP for four RPN detectors (QP values {25,35,47,55}). As expected, for all detectors, miss rate increases with QP. If we examine the RPN trained using the compressed sequence at QP = 55, we observe that for low QP, the performance deteriorates while for high QP the method performs significantly better. The RPN detector trained using the training_PD compressed sequence at QP = 55 learns detection features that are robust to compression, and when evaluated using the training_QE QP = 55 video sequence the miss rate is almost 20% less than the detector trained with original images (Figure 1). Out of the 18 RPN detectors, we select the four detectors shown in Figure 2 because they perform the best over a specific QP interval. As we

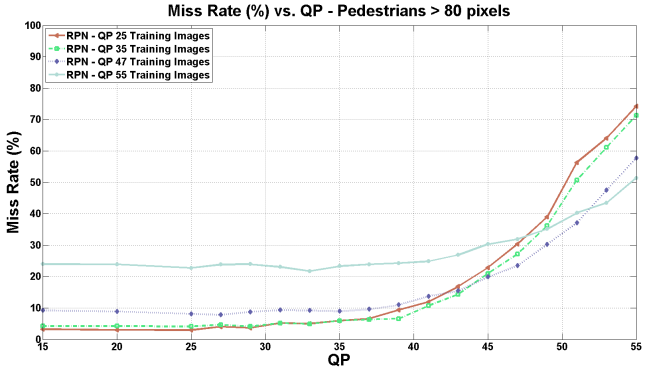

Fig. 2. RPN detectors trained using QP {25,35,47,55}

observed above, degradation in detection performance also depends on what is being sought in a video sequence. When the goal is to detect pedestrians L50, the QP intervals where the detectors perform better change considerably. In this case, we use three detectors with QP {20,35,47}; the results are shown in Figure 3.

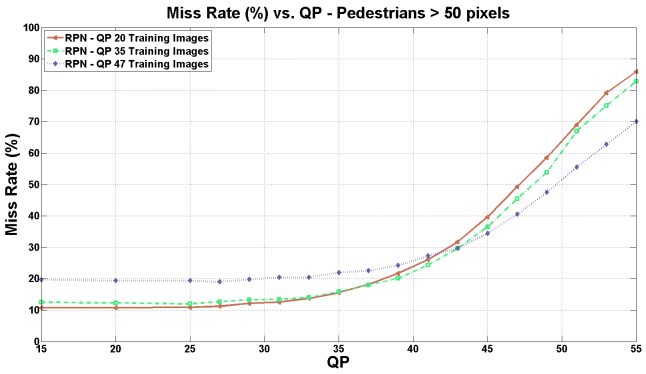

Fig. 3. RPN detectors trained using QP {20,35,47}

3.2. Proposed Two-Stage CNN

Our proposed approach is motivated by the fact that some RPN detectors are more robust to compression than others. Based on this, a system can improve the overall performance by switching between multiple RPN detectors according to the input video compression level. However, in practical applications, transcoding might happen at various points in the network and the final video data-rate is not indicative of the video quality. Therefore, it becomes necessary to estimate video quality for pedestrian detection.

We propose to use a quality estimation component to decide which of the RPN detectors should be used. The overall framework of our proposed system is shown in Figure 4. The system has two components: Quality estimation and a bank of RPN detectors. Quality estimation is the first stage and is based on the Inception-v3 CNN [33]. Inception-v3 classifies each frame of the video into a class associated with an RPN detector and based on the mode of the k consecutive frames makes the decision. If the goal is to detect pedestrians L80 or L50, quality estimation decisions change according to the best performing RPN detector using the results in Figures 2 and 3, respectively. The second stage of our proposed system is a bank of 5 RPN detectors. Each of the detectors is trained using images from a video sequence compressed at a unique quantization parameter. Our proposed system is able to estimate the input video quality and decide on the most suitable detector. This results in an overall reduction in the miss rate value.

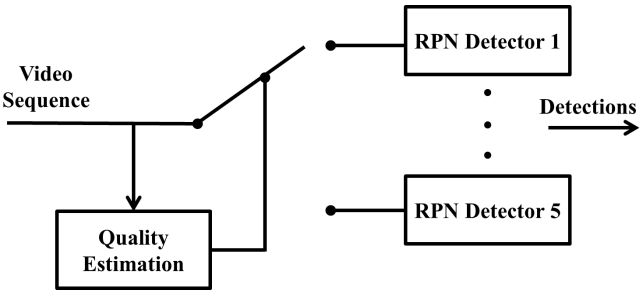

Fig. 4. Proposed Two-Stage Neural Network Block Diagram

3.3. Quality Estimation

Inception-v3 was developed by Google Research. It is 48-layers deep and follows generic design principles based on large-scale experimentation with various architectural design aspects. These guidelines led Google to the development of the current Inception modules, which can be seen as the building block inside the bigger Inception-v3 network. The Inception module uses various convolution filter sizes at each layer, and it computes each convolution in parallel and concatenates the resulting activation maps before passing it to the next layer. When several of these modules are connected, each of the convolutions' activation maps is passed through the mixture of convolutions of the next layer. The use of multiple Inception modules allows Inception-v3 to use parameter weights that favor bigger filter sizes followed by smaller ones, or the other way around depending on the cost function values. This architecture allows Inception-v3 to be sensitive to both local features through smaller convolutions and high-level features through larger convolutions. Inception-v3 achieves 5.64% top-5 error rate on the validation set of the ImageNet entire image ILSVRC 2012 classification task. Furthermore, in the 2015 ImageNet challenge, an ensemble of 4 of these models achieved the second place in the image classification task [33].

For training and validation, we use the transcoded versions of the Caltech training_QE dataset. Labels (or classes) correspond to RPN detectors. For example, the class denoted by RPN-QP35 refers to the RPN detector trained using images from the transcoded video sequence at QP = 35. Images from each transcoded version of the training_QE dataset are assigned the label corresponding to the best performing RPN detector following the results shown in Figures 2 and 3. Two separate CNNs are trained based on each detection goal: pedestrians L80 and pedestrians L50. Figure 5 details the class assignments for each use case. The Inception-v3 quality estimation network was trained using the Caffe deep learning framework on a desktop machine with 3 NVIDIA Titan X GPU cards. The images were randomly shuffled; 90% were used for training (1,106,816 images) and 10% for validation (61,490 images). We used stochastic gradient descent with batch size 32 for 10 epochs, a momentum of 0.9, weight decay of 0.0002 and a fixed learning rate of 0.001 for the initial 5 epochs, and the remaining epochs were done with a fixed learning rate of 0.0001.

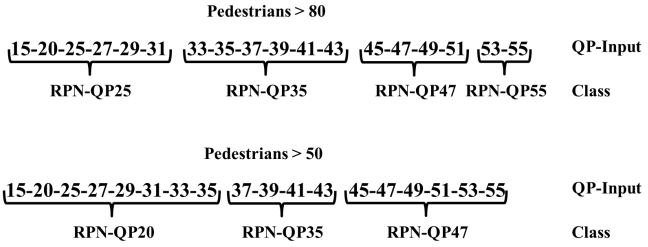

Fig. 5. Class assignments for training the Inception-v3 quality estimation network

4. EXPERIMENTS AND RESULTS

In this section, we evaluate our proposed method, its individual components, and compare 3 approaches each for detecting pedestrians L80 and pedestrians L50. We start with the quality estimation component of our proposed approach. We conduct an experiment using transcoded versions of the Caltech testing_RD dataset at 3 frames/second. A total of $95{,}652$ images are classified into one of the classes detailed in Figure 5. When error rates are calculated on a frame by frame basis ($k = 1$), classification error rates are 5.61% and 10.41% for pedestrians L50 and pedestrians L80, respectively. The error rate is higher for pedestrians L80 due to the existence of four classes compared to three classes for the case of pedestrians L50. Since the quality estimation decision is based on the mode of a temporal window of k samples, we also find the classification error rates for various window sizes. Figure 6 depicts this relationship for pedestrians L80 and pedestrians L50. As expected, classification error rate decreases with the window size k. In our end-to-end evaluation we use $k = 9$, meaning that the quality estimation CNN decides on which RPN detector to use based on samples from 3 seconds.

Quality estimation classification errors refer to deciding on using a suboptimal RPN detector. However, the increase in miss rate introduced by using a suboptimal RPN detector varies according to the particular RPN detector. As can be seen by looking at Figure 2, when detecting pedestrians L80, a frame compressed at QP = 55 should be

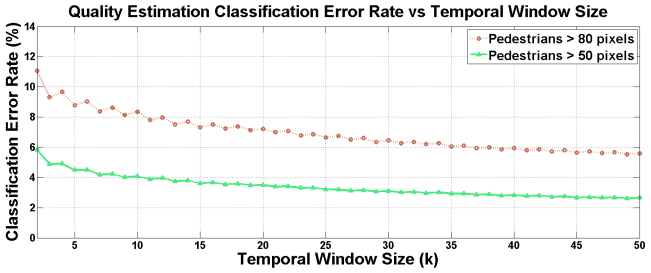

Fig. 6. Classification error rate as a function of the temporal window size k

processed by RPN-QP55, whereas any of the three other detectors yield suboptimal performance. In particular, the use of RPN-QP25 results in the highest increase in miss rate, whereas RPN-QP47 results in the lowest increase. All the errors encountered in our experiment are of the latter type.

In the second experiment, we investigate how much we can improve the overall performance by switching between multiple RPN detectors according to the input video compression level. We assume that we have knowledge of the QP value of the input video and an RPN detector is selected according to Figure 5. We use each transcoded version of the Caltech testing_RD dataset. For each test, we find the miss rate considering pedestrians L50 and pedestrians L80. We also conduct the same tests using an RPN detector trained using original images and an RPN detector trained using images from a mixture of transcoded video sequences. In Figure 7(a) and 7(b), the miss rate is plotted as a function of the QP value for the three mentioned detectors. Our proposed method achieves the lowest miss rate for any transcoded testing video sequence.

Lastly, we evaluate the entire system by emulating adaptive data-rate streaming using the Caltech testing_RD dataset. The process is repeated five times resulting in five consecutive copies of the dataset. To emulate adaptive data-rate streaming, for every temporal window of 10 seconds, one of the 18 transcoded versions of the dataset is selected using a uniform random distribution. On the analytics side, three detectors are used: our proposed quality-adaptive approach, an RPN detector trained using original images and an RPN detector trained using images from a mixture of transcoded video sequences. When the goal is to detect pedestrians L80, the miss rates were: 11.69%, 17.13% and 15.46%, respectively. When the goal is to detect pedestrians L50, the miss rates were: 30.22%, 34.25% and 33.99%, respectively. Our proposed approach achieves the lowest miss rates in both use cases. In the same experiment, if our system is configured to detect pedestrians L50 and evaluated assuming the goal is to detect pedestrians L80 the miss rate increases from 11.69% to 13.10%. Also, if it is configured to detect pedestrians L80 and evaluated assuming the goal is to detect pedestrians L50 the miss rate increases from 30.22% to 31.82%. This demonstrates that the dependency on the use case in our quality-adaptive approach is fundamental to achieving the lowest miss rate.

5. CONCLUSION

In this paper, we proposed a two-stage quality-adaptive convolutional neural network to address the problem of a changing video data-rate. We characterized the impact of video compression on a state-of-the-art pedestrian detector and investigated the use of compressed images for training. Our experimental results demonstrated

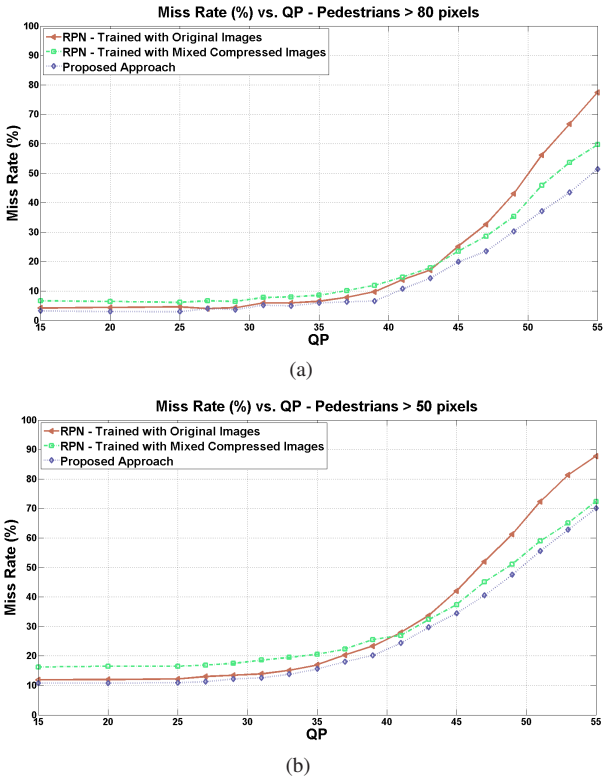

Fig. 7. Bank of RPN detectors performance

that when adaptive data-rate streaming is used our proposed quality-adaptive approach reduces the miss rate compared to the baseline detector.

6. REFERENCES

[1] "Cisco visual networking index: Forecast and methodology, 2015/2020," *Cisco Systems Inc.*, April 2016.

[2] "Fog computing and the Internet of things: Extend the cloud to where the things are," *Cisco Systems Inc.*, April 2015.

[3] 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'.

[4] 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.

[5] K. Yang, E. Y. Du, E. J. Delp, P. Jiang, F. Jiang, Y. Chen, R. Sherony, and H. Takahashi, "An extreme learning machine-based pedestrian detection method," *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 1404–1409, June 2013, Gold Coast, Australia.

[6] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, June 2016.

[7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, August 2013.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] J. Ribera, K. Tahboub, and E. J. Delp, "Automated crowd flow estimation enhanced by crowdsourcing," *Proceedings of the IEEE National Aerospace and Electronics Conference*, pp. 174–179, June 2014, Dayton, OH.

[16] K. Tahboub, N. Gadgil, J. Ribera, B. Delgado, and E. J. Delp, "An intelligent crowdsourcing system for forensic analysis of surveillance video," *Proceedings of the IS&T/SPIE Conference on Video Surveillance and Transportation Imaging Applications*, pp. 94 070I–1–9, February 2015, San Francisco, CA.

[17] A. Redondi, L. Baroffio, M. Cesana, and M. Tagliasacchi, "Compress-then-analyze vs. analyze-then-compress: Two paradigms for image analysis in visual sensor networks," *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, pp. 278–282, November 2013, Sardinia, Italy.

[18] 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.

[19] 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.

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

[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] 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.

[24] 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.

[25] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 794–801, June 2009, Miami Beach, FL.

[26] 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.

[27] 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.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proceedings of the International Conference on Learning Representations (arXiv:1409.1556)*, May 2015, San Diego, CA.

[29] R. Girshick, "Fast R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, December 2015, Santiago, Chile.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Proceedings of the Advances in Neural Information Processing Systems Conference*, pp. 91–99, December 2015, Montréal, Canada.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, April 2015.

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

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *arXiv:1512.00567*, December 2015.