https://doi.org/10.2352/ISSN.2470-1173.2021.17.AVM-174

Boosting computer vision performance by enhancing camera ISP

Peter van Beek, Chyuan-Tyng (Roger) Wu, Baishali Chaudhury, and Thomas R. Gardos
Imaging and Camera Technology Group, Intel Corp., Santa Clara, CA, USA

Abstract

Traditional image signal processors (ISPs) are primarily designed and optimized to improve the image quality perceived by humans. However, optimal perceptual image quality does not always translate into optimal performance for computer vision applications. In [1], Wu et al. proposed a set of methods, termed VisionISP, to enhance and optimize the ISP for computer vision purposes. The blocks in VisionISP are simple, content-aware, and trainable using existing machine learning methods.
VisionISP significantly reduces the data transmission and power consumption requirements by reducing image bit-depth and resolution, while mitigating the loss of relevant information. In this paper, we show that VisionISP boosts the performance of subsequent computer vision algorithms in the context of multiple tasks, including object detection, face recognition, and stereo disparity estimation. The results demonstrate the benefits of VisionISP for a variety of computer vision applications, CNN model sizes, and benchmark datasets.

Introduction

Image Signal Processors (ISPs) convert a raw signal acquired from an imaging sensor to a picture suitable for human consumption. Traditionally, ISPs have been tuned for optimal perceptual image quality for human appreciation. Many applications would benefit from an image processing pipeline that optimizes and normalizes the image characteristics for computer vision engines instead.

In practice, it is challenging to transmit high resolution image data with a large bit-depth and high frame rates from a sensor or an ISP to a computer vision engine, given low power and low latency requirements. To reduce the data transmission load, it is common to downscale the images or to lower the frame rate; however, this can lead to a loss of detail and motion information that is of critical importance in automated/autonomous driving applications.

In [1], Wu et al. proposed to repurpose the ISP for computer vision applications, by **a)** tuning existing ISP blocks for optimal denoising for a subsequent computer vision application; and **b)** introducing novel ISP tone mapping and downscaling blocks that allow reduction of the spatial image size and number of bits per pixel while preserving key image detail information. The operations introduced in [1] are content-aware as well as differentiable and can be trained using back-propagation techniques common in deep learning. This processing sequence, which is called VisionISP and shown in Figure 1, preserves the information that helps a subsequent neural network learn discriminative features even after significant data reduction. The proposed blocks are very lightweight image processing blocks that do not add a significant processing burden over and above the ISP. These methods help reduce data transmission requirements, reduce power consumption, and result in an overall better-optimized pipeline for computer vision applications. Moreover, the approach avoids time intensive and potentially biased manual tuning of the camera ISP for computer vision tasks.

Related recent work in this area includes [2,3,4]. Yahiaoui et al. [2], provide an empirical analysis of the impact of ISP parameters on the performance of computer vision, in an automated driving context. Taylor et al. [3] and Mosleh et al. [4] both describe methods for optimization of the ISP for computer vision accuracy. Both propose to use general non-linear optimization techniques capable of tuning tens or hundreds of ISP parameters in a black-box setting. The approach described in Wu et al. [1] is different in that novel machine learning-based image processing blocks are proposed that allow the use of efficient gradient-based end-to-end optimization techniques.

In Wu et al. [1], experimental results were included showing the benefit of VisionISP in the context of object detection. In this paper, we present new experimental results across a wider variety of computer vision tasks, multiple CNN backbone models, and a wider range of usage configurations. Specifically, we demonstrate significant improvements in accuracy when using VisionISP in object detection, face recognition, and stereo disparity estimation, and across a wide range of data reduction factors and model sizes.

In the next section, we recap the VisionISP blocks as introduced in [1], discuss the automated training of VisionISP, and provide some example image outputs. The subsequent section describes our experimental results in detail.

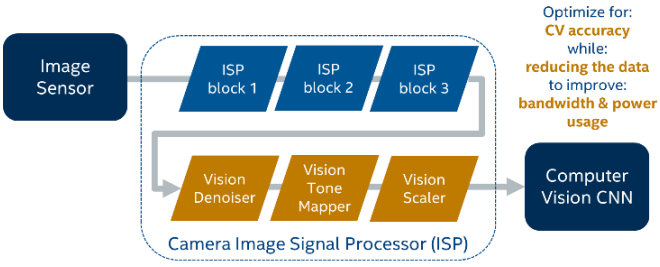


Figure 1: VisionISP overview.

VisionISP Components and Training

In the following, we will discuss the main VisionISP blocks: **a)** Vision Denoiser, **b)** Vision Scaler, and **c)** Vision Tone Mapper.

Vision Denoiser

The presence of strong noise in camera image data has a tendency to reduce the accuracy of computer vision algorithms. On the other hand, application of heavy noise reduction in practice results in a loss of detail information, and such detail may be critical for the specific application. Hence, properly tuning the denoiser in the context of a computer vision task is important, and

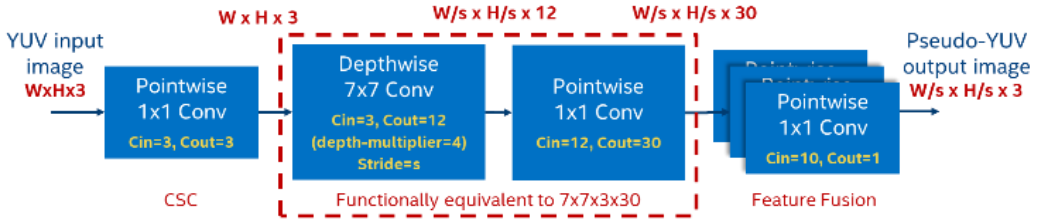


Figure 2: Trainable Vision Scaler (TVS).

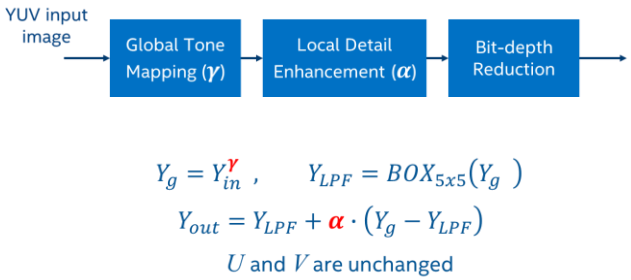


Figure 3: Vision Tone Mapper (VTM).

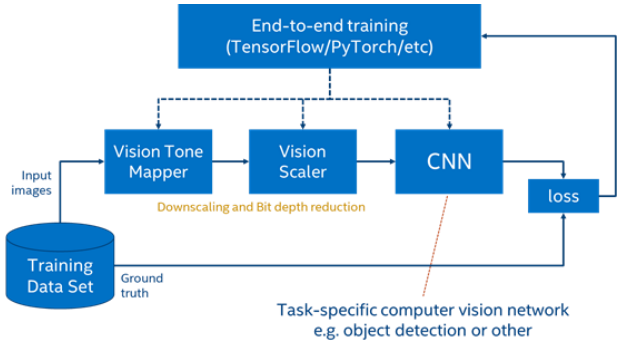


Figure 4: Training VisionISP blocks in the context of a computer vision task, where a CNN implements the computer vision task.

such vision-optimized tuning likely has a different result from tuning to optimize the image quality for human viewers.

The vision denoiser consists of a combination of spatial and motion-compensated temporal filtering blocks. We use a denoiser from an existing ISP without algorithm or hardware modifications. It has several parameters that can be tuned automatically using the optimization approach described in [5]. To optimize the denoiser specifically for computer vision tasks, the objective function is defined as the L1 norm of the error between CNN feature maps resulting from the denoised image and reference CNN feature maps. The reference CNN feature maps originate from a set of noiseless reference images. Please see [1] and [5] for further detail.

Vision Tone Mapper

The vision tone mapper (**VTM**), illustrated in Figure 3, consists of a global tone mapping operator followed by a local detail boosting operator, and finally, bit-depth reduction. The global tone mapping operator is defined by a non-linear function, which can be any differentiable function having trainable parameters. Here we use a simple gamma function, with a single trainable parameter $\boldsymbol{\gamma}$. Local detail enhancement is performed by a standard unsharp masking technique, using a fixed 5x5 box-filter and a trainable parameter $\boldsymbol{\alpha}$ to adjust the enhancement strength. Bit-depth reduction is a uniform quantization step and can be performed simply by bit-shifting. In practice, the bit-depth reduction step is performed after the vision scaler.

Vision Scaler

The trainable vision scaler (**TVS**) is illustrated in Figure 2. As described in [1], TVS is a small trainable neural net consisting of a color space conversion (CSC) layer, a feature extraction layer, and a feature fusion layer. The CSC layer is a 1x1x3x3 pointwise convolutional neural net layer. The feature extraction layer is equivalent to a 7x7x3x30 convolutional layer, but is decomposed into a depthwise convolution and pointwise convolution for computational efficiency. Downscaling is controlled by the striding in the depthwise convolution step. Note that non-integer scaling factors can be achieved by non-uniform striding. The final step fuses the feature channels back into 3 channels, as expected by computer vision algorithms, and is implemented as a 1x1x30x3 pointwise convolution with grouping of 3, or equivalently, three 1x1x10x1 pointwise convolutions.

VisionISP Training

The vision tone mapper and vision scaler can be trained using machine learning frameworks such as TensorFlow/Keras/PyTorch, as illustrated in Figure 4. These blocks can be simulated using trainable layers and pre-pended to any existing computer vision model that ingests 3-channel images. Subsequently, these layers can be trained together with the computer vision model in an end-to-end manner, using a task-specific loss.

In addition, ML frameworks allow training only the computer vision model by itself, while freezing the parameters of the VisionISP layers, and vice-versa. Such training regimes can be useful in practice, to avoid re-training of either the VisionISP layers or the computer vision model layers, as might be desired.

In Figure 5, we show a few examples of output images of VisionISP, where TVS and VTM were trained end-to-end with an existing object detection model, and processing included downsampling and bit-depth reduction. The effect of the trainable color space conversion is obvious, while the effects of global and local tone mapping can also be seen. Although the images appear quite different from natural images for human viewing, the accuracy of the subsequent computer vision model improves, as will be shown in the next section.

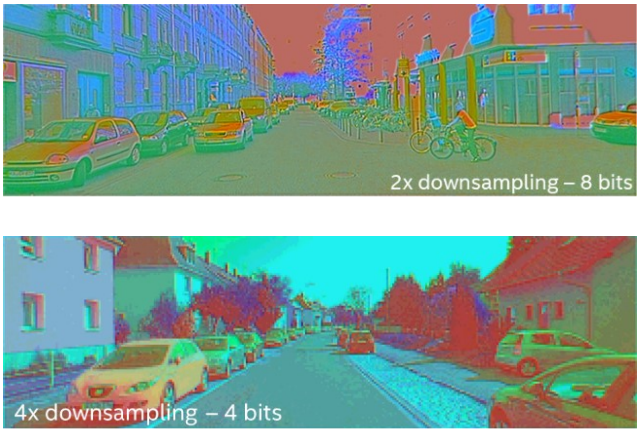


Figure 5: Example VisionISP output images.

Experimental Results

We report our results on several computer vision tasks: object detection, face authentication/recognition, and stereo disparity estimation. In our experiments, we have used well-known computer vision networks, as well as publicly available datasets and benchmarks, as detailed below.

We compare our accuracy results to baseline results without VisionISP. In the baseline configuration, bilinear image downscaling was used instead of TVS. In the baseline configuration, the global tone mapping and local enhancement blocks of VTM were simply omitted, prior to bit depth reduction. In most cases, we report the results of end-to-end training with VisionISP. At the end of this section, we also report results for a training setup where VisionISP and the computer vision model are trained separately.

Computer vision tasks and data sets

For object detection, we have used the open-source TensorFlow implementation of the SqueezeDet+ network [6], with a variant of SqueezeNet as its backbone. We used the KITTI data set [14] for training and evaluation, following the standard training and evaluation procedures and accuracy metrics. The loss function and other training settings were adopted from [6].

For object detection, we have also used our own TensorFlow-Keras network implementation with a MobileNetV2 backbone [7] and a YOLO detection head [8]. With this pipeline, we report results both on the KITTI data set, and on the 'daytime' image subset of Berkeley Deep Drive (BDD100K) data set [9].

For face recognition/authentication, we have used an open-source implementation of FaceNet [10] available from [11]. The FaceNet model uses either Inception-ResNet-v1 or SqueezeNet as its backbone. The VGGFace2 data set [15] was used for training, and the Labeled Faces in the Wild (LFW) data set [16] was used for evaluation. Training used the standard softmax cross-entropy loss and further training settings adopted from [11].

For stereo disparity estimation, we have used the MADNet [12] and DispNet [13] networks, using the open-source implementation available from [12]. Both networks regress dense disparity maps using UNet-like network architectures. The Middlebury 2014 data set [17] was used for training and evaluation, using an 80%-20% train/test split, and training settings adopted from [12].

Results demonstrating the benefits of the proposed enhanced ISP are shown in the next subsections.

Object Detection

In Figure 6, we show the mean average precision (mAP) for object detection with and without the use of the vision tone mapper (VTM) and vision scaler (TVS), for various networks and data sets. Each graph shows mAP for various combinations of bit precision (8/4/2) and downscaling factor (1x, 2x, 3x, 4x). The reference mAP obtained without downscaling and with 8 bit precision is shown on the far left, if available. The results show that a higher mAP can be achieved with VisionISP, compared to not using VisionISP. A relative improvement in mAP up to ~26% can be achieved when large downscaling factors are applied, while smaller improvements are obtained for small downscaling factors.

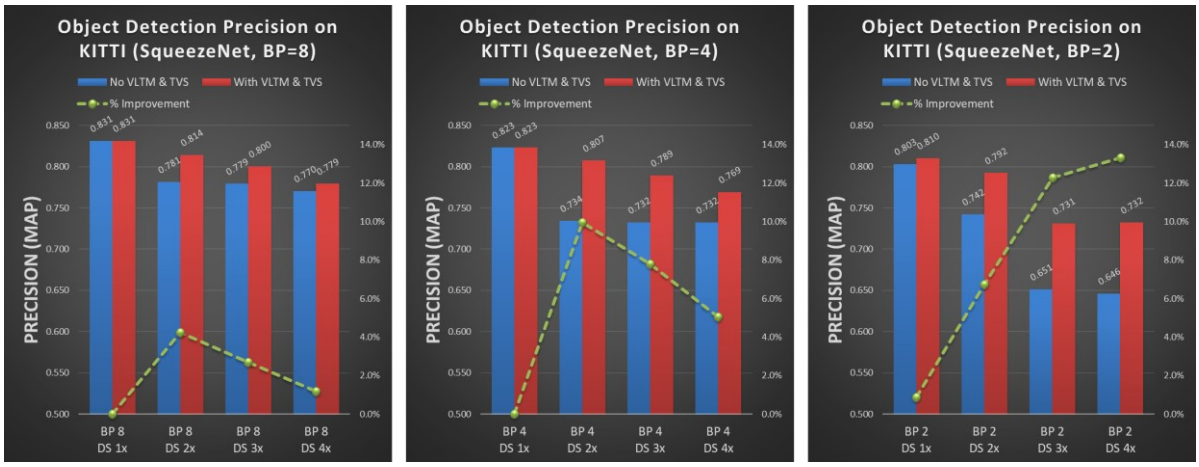
Furthermore, as shown in Figure 6.a, in some cases it is possible for VisionISP to achieve significant data reduction (e.g. 2x downscaling on each axis as well as bit depth reduction from 8 to 4) with very little degradation in mAP compared to the reference case at full resolution and full bit depth.

Face authentication

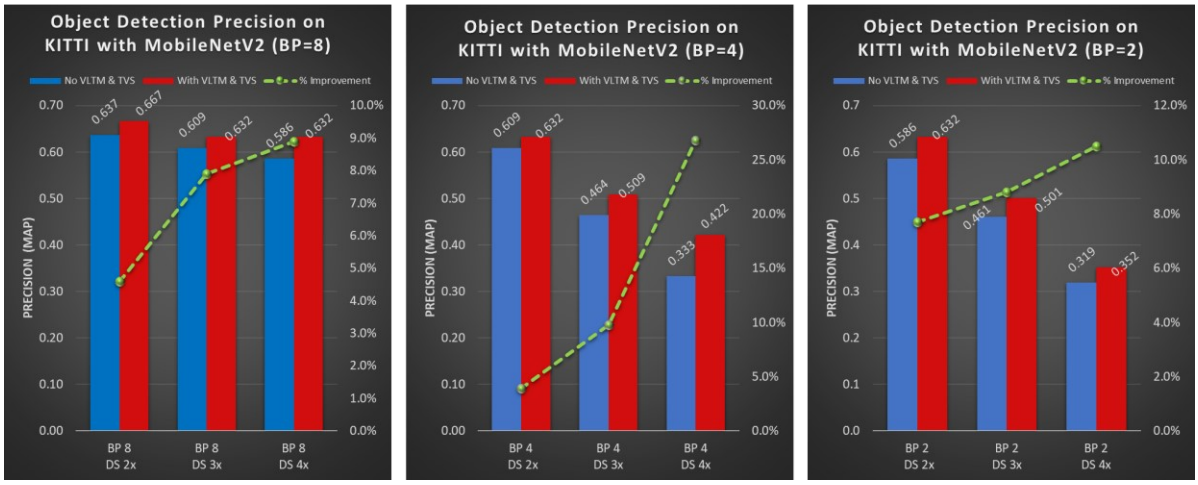
In Figure 7, we show the face validation rates with and without the use of the vision tone mapper (VTM) and vision scaler (TVS), with two network backbones: Inception-ResNet and SqueezeNet. The validation rate is the True Acceptance Rate at a False Acceptance Rate of 0.001. Each chart again shows results for various combinations of downscaling factor and bit depth reduction. The results for both networks show that a higher accuracy can be achieved with VisionISP, compared to not using VisionISP. In particular, when reducing bit precision from 8 to 2, the validation rate is significantly better with VisionISP, compared to the baseline.

The relative improvement in accuracy is higher with SqueezeNet, a small backbone of about 1M parameters, than with Inception-ResNet, a very large backbone of about 50M parameters. Two aspects may explain this: the first is that the baseline accuracy with Inception-ResNet is higher in absolute terms and there is less room for improvement, and the second is that the impact of the additional trainable layers in VisionISP may be significantly more helpful for the smaller SqueezeNet backbone.

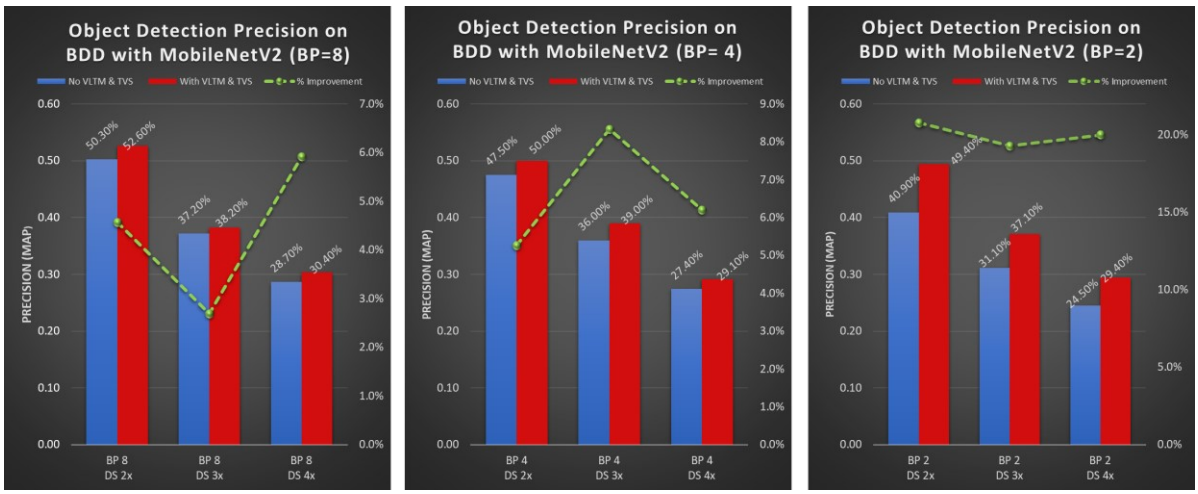
Also note that, with VisionISP, it is possible to achieve a validation rate that is the same or even better than the baseline at 8 bits and no downscaling, even with strong downscaling and low bit precision.



a) mAP for SqueezeNet with SqueezeDet head on KITTI data set

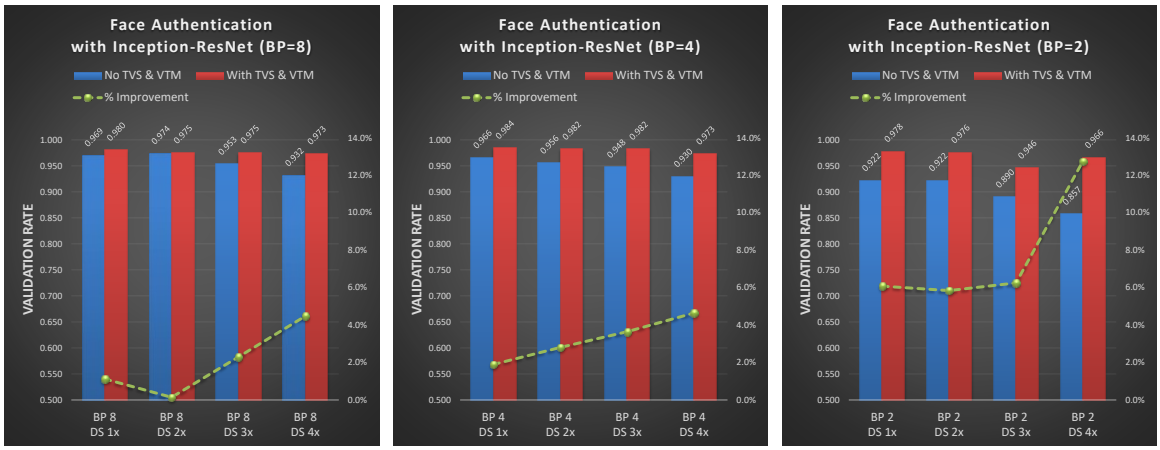


b) mAP for MobileNetV2 with YOLOv2 head on KITTI data set



c) mAP for MobileNetV2 with YOLOv2 head on BDD100K data set

Figure 6: Object detection accuracy with and without the proposed vision scaler and vision tone mapper (higher is better). Red and blue bars show the mAP with and without TVS and VTM, and the green points show relative improvement in mAP.



a) Validation rates for FaceNet with Inception-ResNet backbone trained on VGGFace2 and evaluated on LFW

b) Validation rates for FaceNet with SqueezeNet backbone trained on VGGFace2 and evaluated on LFW

Figure 7: Face authentication accuracy with and without the proposed vision scaler and vision tone mapper (higher is better). Red and blue bars show the validation rate with and without TVS and VTM, and the green points show relative improvement in validation rate.

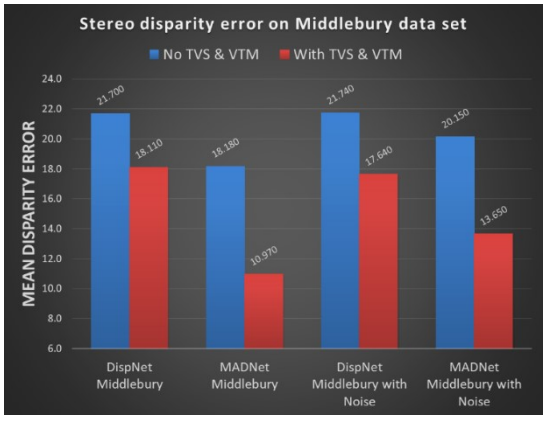


Figure 8: Stereo disparity mean error on the Middlebury 2014 data set with and without the proposed vision scaler and vision tone mapper (lower is better).

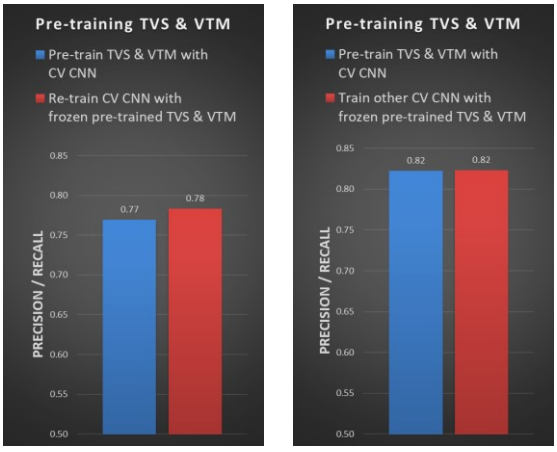


Figure 9: Accuracy is retained when training an object detection network with pre-trained frozen TVS & VTM (instead of end-to-end training). On the left, the object detection backbone is the same during pre-training; on the right, the backbone is different.

Stereo disparity estimation

In Figure 8, we show the mean disparity error with and without the use of the vision denoiser, vision tone mapper (VTM) and vision scaler (TVS), for stereo disparity estimation. We include results with a DispNet network backbone and a MADNet backbone. Also, we show results with the original images as well as an image data set where artificial noise was added. The artificial noise was modeled to resemble sensor noise at a analog gain of 16x, and an SNR of approximately 15.8 dB. The vision denoiser was used in the case of the noisy data. In this experiment, downscaling was set to 8x and the bit precision was 2. The results show that TVS and VTM improve the disparity estimation accuracy (lower disparity errors).

End-to-end training versus pre-trained VisionISP

As mentioned, it is possible to train only the computer vision model, using pre-trained VisionISP layers. In this section, we compare the accuracy results of the following:

a) Training TVS/VTM and computer vision model end-to-end in a single training session,
b) Use the resulting TVS and VTM weights from a), then train the computer vision model with frozen TVS/VTM weights.

In Figure 9, on the left, we show results for the case where the computer vision model was SqueezeNet in both steps a) and b). On the right, we show results for the case where, a SqueezeNet backbone was used in step a) but an Inception-ResNet backbone was used in step b).

In both cases, the accuracy obtained by re-training a computer vision model with frozen TVS/VTM weights is about the same as the accuracy for end-to-end training. Furthermore, this is the case even when using another backbone for the computer vision task as in the second case. This means that the TVS/VTM trained weights generalize to other backbones, and that TVS/VTM can be pre-trained without access to the backbone chosen by users. This is relevant for practical deployment of these techniques, since it enables providing sets of pre-trained TVS/VTM weights to users, allowing the users to integrate TVS/VTM while only needing to train the main computer vision model.

Conclusions

In this paper, we built upon the work described in [1]. We propose to enhance the camera ISP with a few lightweight, machine-trainable, content-aware image processing techniques for the purpose of optimizing accuracy of computer vision tasks.

In particular, we evaluated the proposed VisionISP concepts across a variety of computer vision tasks, network backbones, and data sets. Our evaluation included object detection, face authentication, and stereo disparity estimation. Our evaluation results demonstrate that the proposed techniques can significantly improve computer vision accuracy when downscaling and/or bit depth reduction is required to reduce the data rate or image resolution. Also, in some cases, these techniques enable downscaling and bit depth reduction while retaining computer vision accuracy; hence significantly reducing power consumption and bandwidth usage in embedded applications.

References

[1] C. Wu, L. Isikdogan, S. Rao, B. Nayak, T. Gerasimow, A. Sutic, L. Ain-kedem and G. Michael, "VisionISP: Repurposing the image signal processor for computer vision applications," IEEE Int. Conf. on Image Processing (ICIP) 2019.

[2] L. Yahiaoui, J. Horgan, B. Deegan, S. Yogamani, C. Hughes, and P. Denny, "Overview and Empirical Analysis of ISP Parameter Tuning for Visual Perception in Autonomous Driving," Journal of Electronic Imaging, Sept. 2019.

[3] D. Taylor, A. Sharma, K. St. Arnaud, and D. Tokic, "Automated optimization of ISP hyperparameters to improve computer vision accuracy," IS&T Electronic Imaging (AVM), 2020.

[4] A. Mosleh, A. Sharma, E. Onzon, F. Mannan, N. Robidoux and F. Heide, "Hardware-in-the-loop End-to-end Optimization of Camera Image Processing Pipelines," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2020.

[5] J. Nishimura, T. Gerasimow, S. Rao, A. Sutic, C. Wu, and G. Michael, "Automatic ISP image quality tuning using nonlinear optimization," IEEE Int. Conf. on Image Processing (ICIP) 2018.

[6] B. Wu, A. Wan, F. Iandola, P. Jin, and K. Keutzer, "SqueezeDet: Unified, small, low-power, fully convolutional neural networks for object detection for autonomous driving," IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW) 2017.

[7] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2018.

[8] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," https://arxiv.org/abs/1612.08242, 2016.

[9] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2020.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015.

[11] D. Sandberg, "Face Recognition using TensorFlow," https://github.com/davidsandberg/facenet, 2018.

[12] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. di Stefano, "Real-time self-adaptive deep stereo," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2019.

[13] N. Mayer, E. Ilg, P. Hauser, P. Fischer, D. Cremers, A. Dosovitskiy and T. Brox, "A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2016.

[14] A. Geiger, P. Lenz and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2012.

[15] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," IEEE Conf. on Automatic Face and Gesture Recognition (F&G), 2018.

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," University of Massachusetts, Amherst, Technical Report 07-49, October 2007.

[17] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," German Conference on Pattern Recognition (GCPR 2014), Münster, Germany, September 2014.

Author Biographies

Peter van Beek is a senior algorithms engineer with the Intel Imaging and Camera Technologies Group. He is supporting development of deep learning inference engines as well as optimization of camera imaging pipelines. Previously, he was with the Intel Autonomous Driving Group and Mobileye. Before joining Intel, Peter was a technical lead at Sharp Laboratories of America. He is a senior member of the IEEE and received a PhD from Delft University of Technology.

Chyuan-Tyng Wu received a Ph.D. in Electrical and Computer Engineering from Purdue University. His research was focused on computer vision and a depth information capture system. After college, he is working as an Algorithm Engineer in Imaging and Camera Technologies Group at Intel. His work includes multiple fixed function developments for the imaging signal processor and emerging IP incubation for computational photography and the machine learning applications.

Baishali Chaudhury is a Machine Learning Engineer in the Advanced Solutions group in the Imaging and Camera Technologies Group at Intel. Her work centers around developing deep learning training and inference pipelines for use cases such as object detection and face detection, steered towards the development of imaging signal processors. Previously, Baishali worked as a software engineer in the Data Platforms group at Intel, to optimize deep learning networks for various Intel edge accelerators. She received a PhD in Computer Science and Engineering from the University of South Florida.

Thomas Gardos is a Principal Engineer and manager of an imaging and computer vision R&D team at Intel. Previously, he managed Intel's video compression algorithm team and represented Intel in video standards bodies. Before joining Intel, Tom was in the Imaging Research Labs at Eastman Kodak and a graduate of Kodak's 2-year Imaging Scientist training program. Tom received his MS and PhD in Digital Image and Signal Processing from Georgia Tech and his BSEE from University of Delaware.

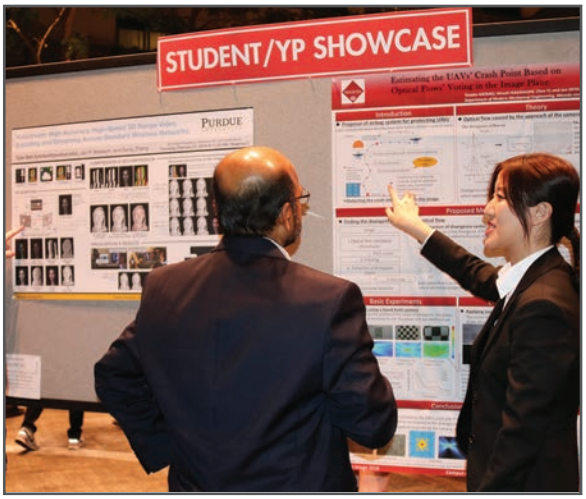
JOIN US AT THE NEXT EI!

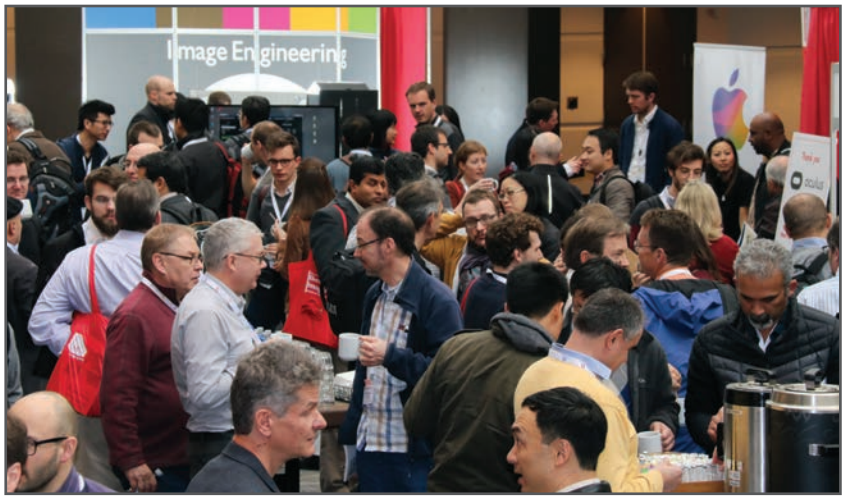
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org