https://doi.org/10.2352/ISSN.2470-1173.2019.15.AVM-044

Optimization of ISP parameters for Object Detection algorithms

Lucie Yahiaoui, Ciarán Hughes, Jonathan Horgan, Brian Deegan, Patrick Denny and Senthil Yogamani
Valeo Vision Systems, Tuam, Co.Galway, Ireland

Abstract

In autonomous driving applications, cameras are a vital sensor as they can provide structural, semantic and navigational information about the environment of the vehicle. While image quality is a concept well understood for human viewing applications, its definition for computer vision is not well defined. This gives rise to the fact that, for systems in which human viewing and computer vision are both outputs of one video stream, historically the subjective experience for human viewing dominates over computer vision performance when it comes to tuning the image signal processor. However, the rise in prominence of autonomous driving and computer vision brings to the fore research in the area of the impact of image quality in camera-based applications. In this paper, we provide results quantifying the accuracy impact of sharpening and contrast on two image feature registration algorithms and pedestrian detection. We obtain encouraging results to illustrate the merits of tuning image signal processor parameters for vision algorithms.

Introduction

Perceptual image and video quality for human viewing applications is now becoming the subject of standardization [1], taking into account key concepts such as perceptual quality, faithfulness and naturalness, and driven by the development of compression algorithms in particular for multimedia content. In contrast, the definition of image quality for computer vision is not yet as clearly defined, although attempts are being made to standardize [2]. In the automotive industry, cameras are used for two distinct application classes: human vision (HV), whereby scene information from the environment of the vehicle is displayed to a driver or other vehicle occupant (e.g., rear view and multi-camera surround view monitoring) and computer vision (CV) for advanced driver assistance systems (ADAS) [3, 4]. With the move toward autonomous driving platforms, coupled with significant advances in machine learning in the last decade, computer vision performance is expected to make significant jumps over the previous systems as cameras that were previously used to help the driver are now becoming a key sensor for near field sensing. This interest in image quality for computer vision is reinforced by the fact that the priority in terms of the automotive system is swinging away from traditional viewing aspects towards maximizing the performance of computer vision algorithms, given the growing importance of automated driving.

The Image Signal Processor (ISP) is a processing block that converts a raw digital image into a usable image for a given application (e.g. a colour image for viewing). This conversion is quite complex and includes a number of discrete processing blocks that can be arranged in a different order depending on the ISP (Figure 1). It also typically includes some control logic for the analog components of image capture, such as auto exposure and gain control (AEC and AGC, respectively) and feedback loops, such as auto white balance (AWB). Each ISP has its own unique features, but almost all ISPs have the same basic blocks and processing flow.

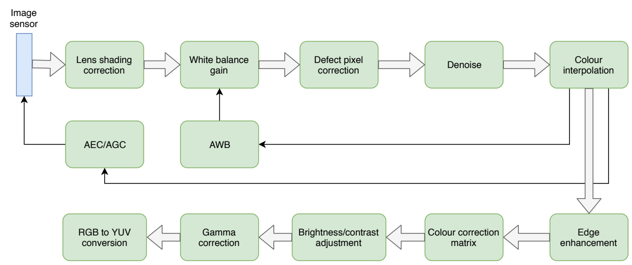

Figure 1. *Example of an ISP pipeline*

In systems that have multiple applications from a single set of sensors, such as surround view monitoring and near field sensing, Human Vision (HV) and Computer Vision (CV) functions have to share a single ISP pipeline that is created traditionally according to the human visual quality of the surround view system. The parameters of this ISP can be modified during the project in order to improve the visual quality to meet requirements from the customer. All algorithms could be impacted by, for example, SNR (Signal-to-Noise Ratio) degradation introduced by changes. Geometric vision algorithms are inherently sensitive to ISP preprocessing changes as the pixel level operations such as feature extraction generally rely on fixed statically tuned kernel sizes and parameters as well as fixed saliency thresholds. Regarding machine learning algorithms, they may be somewhat more robust as long as their model is trained with a high variability of training samples.

In the next section, we will describe the test setup we used to analyze the potential impact of ISP on CV performance. Due to data availability (unavailability of image data without ISP already applied) and difficulty in modeling complex physical processes, we cannot present *absolute* measures of performance with various ISP settings. However, instead, we aim to simply answer the question of *whether there is an impact or not*, by using some well known image filtering techniques to roughly approximate configurations of the ISP chain.

This paper expands significantly on the authors' previous short paper publication [5].

Test setup

The study presented in this paper shows results obtained for sharpening and contrast from a pixel-level processing perspective and shows their impact on pedestrian detection performance, measure by key performance indicators (KPIs) of true and false positive rates. These two blocks are typical ISP processing blocks that more than any other parameters, are driven by subjective experience rather than objective fitness for the application. It's important to note that we are not trying to model the exact method that various ISP suppliers use for sharpening and contrast. This is impossible, as this is generally intellectual property that is not publicly available. Additionally, the property of an image known as "contrast" is a complex interaction of several ISP components, e.g. colour filter pattern, lens shading correction, exposure time, analog/digital gain, black level correction, etc. Rather, we aim to address the broader question of whether contrast adaption and/or artificial image sharpening have an impact on computer vision, regardless of their source.

The study has been performed on images extracted from a video stream captured using a fisheye lens, as these are sensors that are commonly used in systems that have requirements for both human viewing and computer vision (i.e surround view camera systems). Note that these images are not raw, this means that a basic ISP has been applied to prior the tests; this is due to limitation in obtaining data without any ISP applied. These tests will be rerun in the future with raw images to give a better objective understanding of the impact of ISP on CV. Figure 2 is an example of one of the images used in the study.

In the results for the feature registration, we examine the 100 most salient matched points in a patch of the image, looking into the effect of ISP on each feature extractor independently. Over 100 images have been used in the study. In computer vision, a feature is a salient part of an image (point, blob, edge etc.) which reduces the amount of data to be processed, focuses on the relevant parts of the image, may be temporally robust and is further processed by the next stage of the CV algorithm pipeline. The feature descriptors/detectors used in this investigation are invariant to uniform scaling, orientation and illumination [6]: Scale-invariant feature transform (SIFT) [7] and Accelerated-KAZE (AKAZE) [8, 9]. To show the impact on feature extraction, the results obtained for the AKAZE and the SIFT feature detectors are presented in the sharpening and contrast sections (e.g.: Figure 3).

The features are extracted in six frames in total ($n \dots n+5$), and matched between frame n and each of the successive frames in the sequence ($n+1$ to $n+5$).

We also analyze the performance of a light pedestrian detection (PD) algorithm pipeline consisting of an Adaboost candidate generation followed by a small CNN based validation, showing preliminary results of the impact of sharpening and contrast. A catalog of 93 videos has been used for testing. The metric used to quantify the impact is a standard KPI used in the industry: intersection over union (IOU), which measures the difference between the annotated (ground truth) and the detected bounding boxes. The larger the IOU value the better and above a defined threshold per class instance in the image it is considered a true positive.

Image enhancement

In the previous section, we briefly described the computer vision components that we are going to use to form the basis of the tests. Here, we will describe in more detail the image enhancement that we will apply to the images to model the ISP settings. We have already noted that a true model of sometimes unknown ISP components is near impossible, so what we aim to do here is provide some form of approximation. With that in mind, we will in fact look at several algorithms that are designed to achieve the same effect, e.g. for sharpening we will investigate Laplacian and unsharp mask techniques in various configurations.

Figure 2. *Example of fisheye images used in the study. The two red square are showing the ROI selected to show results in the next sections of the paper*

Figure 3. *AKAZE feature detector on a patch of the image*

Sharpening

The human vision system is highly sensitive to edges, and is widely believed to be an important early stage in human visual processing [10]. In images, borders between objects (edges) are perceived because of the intensity change (higher intensity transitions equates to a "sharper" image). Sharpening is widely used to post process images by enhancing the scale of the intensity transitions. That is, increasing the difference between dark and bright regions accentuates edges, and is often considered visually pleasing.

In this paper, two techniques have been used to sharpen images. The first technique is using Laplacian filters to enhance fine details contained in the high-frequency regions (Figure 4 (top)). The kernels are designed to increase the brightness of the center pixel relative to the original ones. The Laplacian operators (L_4 and L_8) are a 2D isometric measure of the second spatial derivative of an image:

$$L_4 = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}, L_8 = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix}$$

The second technique used is the unsharp masking (USM).

It uses an unsharpened (blurred) negative image mask of the original image, combined through a per-pixel weighted sum with the positive original image, to create a sharpened version. The steps are shown in Figure 4 (bottom). Here we use a set of Gaussian kernels with varying sizes (3×3, 9×9 and 19×19) to create the blurred image, and use a weighting of 0.5 on the blurred image.

The two techniques have been applied to the images (Figure 5). It is visually apparent that images sharpened using the Laplacian filters have the noise enhanced and halo artifacts are introduced.

Using the USM technique, some halo artifacts appear in the images, but these artifacts are less pronounced than the Laplacian filters. This is understandable from the techniques applied, as USM is a subtractive rather than additive technique, and is the reason that USM remains a popular technique in digital photography.

Contrast enhancement

In visual perception, contrast is determined by the difference between the brightness and colour of objects within the same field of view. Indeed, it is well understood that luminance contrast plays a larger role in the human visual system than absolute luminance [11]. Histogram equalization is a typical approach for contrast enhancement, and as such we examine two such approaches: the first is the global, but possibly naïve, histogram equalization (Figure 6), and the second is a local approach known as Contrast Limited Adaptive Histogram Equalization (CLAHE) (Figure 7) [12].

These two techniques have been used to examine the impact of contrast on the performance of computer vision. The naïve histogram equalization impacts the global contrast of the image and visually could be said to give better results for images that are over or under-exposed. The advantage of this method is that it is not computationally expensive and it is invertible, except for quantization. As it impact the image globally and without discrimination, the noise can be inadvertently enhanced.

CLAHE computes several histograms for each section of the image where the intensity values are redistributed. The contrast amplification is limited by clipping the histogram at a predefined value before the computation (called clip limit). The clipped part is then redistributed equally among all histogram bins. As this technique is expensive computationally, different interpolations are used to improve the computation without compromising the quality (bilinear and linear interpolations).

The two methods have been applied to our image set. The histogram equalization results in images shifted to a different intensity. The consequence is that some high intensity details (such as clouds, some boundaries, etc.) disappear from the image and others less intense areas become more distinguishable (Figure 8).

Visually, the images generated by the CLAHE method seem to contain more details and show that the dark areas from the original image become darker and the light become lighter, as expected. It can be observed that the noise is enhanced with the size of the tiles and the clip limit applied and halo artifacts are produced along edges.

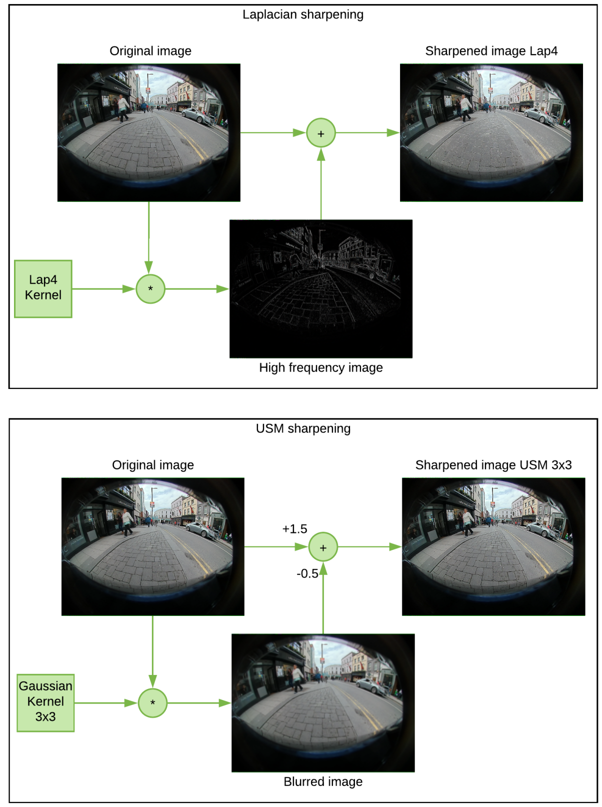

Figure 4. *Laplacian and Unsharp masking techniques*

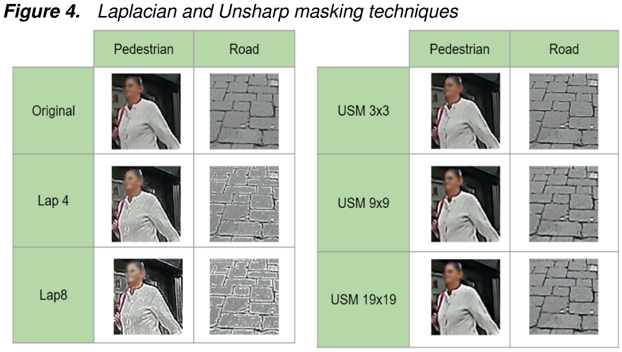

Figure 5. *ROI after Laplacian filter and Unsharp masking. Original: no sharpening, Lap4 & Lap8: Laplacian technique using the kernels containing 4 or 8 in the middle element, USM 3x3 or USM 9x9 or USM 19x19 refer to USM using a Gaussian kernel (size kernel: 3x3 or 9x9 or 19x19)*

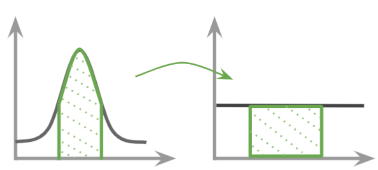

Figure 6. *Histogram equalization diagram*

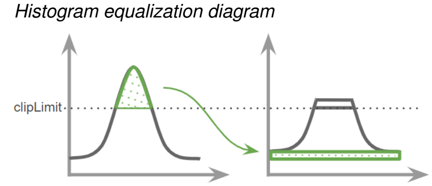

Figure 7. *CLAHE diagram*

Results

Edge Detection

We won't present results on the "goodness" of edge detection, as this is a rather unquantified problem, particularly compared to feature detection where registration performance can be measured. However, examining edges gives a useful visual interpretation of the impact of the different image enhancement techniques, and aids the reader to interpret the further results in a more intuitive way. For the edge detection, we use the very common Canny edge detector with Sobel filter [13] (Figure 9).

With the edge enhanced images, the noise that has been added by the Laplacian filter seems to be detected as edges in some cases. Moreover, halo artifacts produced by the Laplacian filters are detected as edges contrary to the halo artifacts created by USM, as they are not strong enough.

Regarding the edge detection on the contrast enhanced images (Figure 10), visually it can be seen that the image after histogram equalization is really close to the original image, the noise is not detected as an edge in this case. Moreover, some edges disappear in bright areas and appear in dark areas. It can be observed that the halo artifacts produced by the CLAHE method are detected as edges. Every edge seems to be doubled and the noise created is spread across the entire image and being detected as edges.

Feature Registration

As explained earlier in the paper, impact on feature detection has been studied. To do that, a sub-image has been extracted from a *frame n* (e.g.: sub-image containing the car in Figure 3). Keypoints are matched by identifying the nearest neighbours, ratio of closest-distance to second-closest distance. If it is greater than a threshold value, then they are rejected to keep only the good matches. The true positive matching points are called inliers. Two feature detectors (SIFT and AKAZE), have been applied to find inliers between the sub-image extracted and the original image (sharpened or not). To check the continuity of the matching points, the feature detectors have been applied between the sub-image from the frame n with the five next frames (frames $n, n+1, n+2, ...n+5$). The results presented below have been obtained by applying the AKAZE and SIFT feature detectors. Each block of the chart shows the number of inliers found between the sub-image extracted from the frame n and subsequent frames in the video sequence. It can be observed that the true positive (TP) percentages obtained for AKAZE and SIFT decreases when the frame number increase. This is as expected, as there is a greater motion of the scene between frames.

Sharpening

Regarding both tests, the images sharpened using the Laplacian filters never make the TP percentage to increase (Figure 11 & Figure 12). For the frame $n+1$ to $n+5$, it can be noticed that the TP percentages are higher after unsharp masking than with the original in the AKAZE results. Regarding the test made with the SIFT detector, it can be observed that the impact is more mitigated as the unsharp masking makes the TP percentages lower in the majority of cases. The results of this test show that USM method can have a positive impact on performance of the AKAZE feature detector.

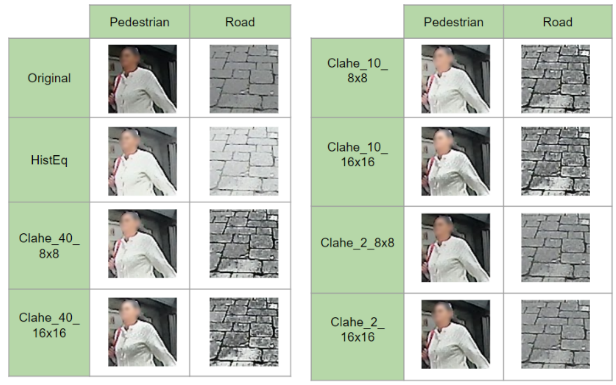

Figure 8. *ROI after Histogram equalization and CLAHE. Original: no contrast, histEq: Histogram equalization, CLAHE_clipLimit_tileSize: clipLimit = 2, 10 or 40 & tileSize = 8x8 or 16x16*

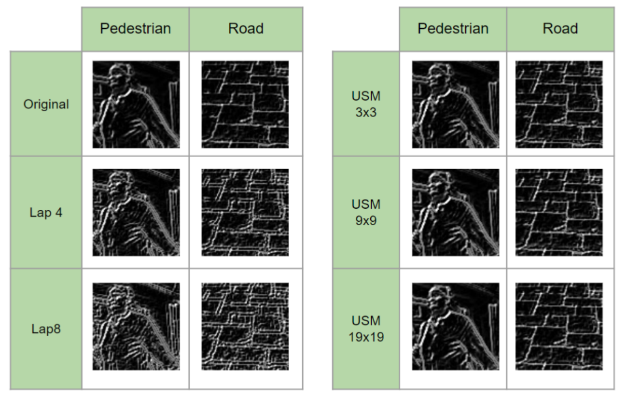

Figure 9. *Edge detection on original and sharpened images*

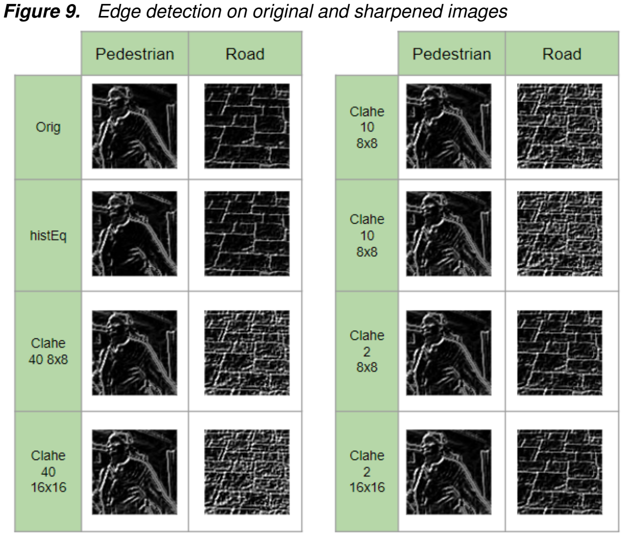

Figure 10. *Edge detection on original and contrast enhanced images*

Contrast enhancement

The same tests as the tests done for sharpening have been made. The results presented below have been obtained by applying the SIFT and the AKAZE feature detectors (Figure 14 & Figure 13).

The contrast enhancement made using histogram equalization technique never causes the TP percentage to increase. In the

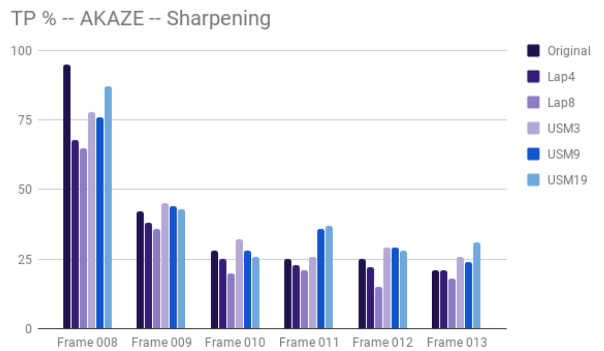

Figure 11. *True positive percentages of AKAZE given by frame and by sharpening technique used. The sub-image used has been extracted from the frame 8 and the tests have been applied from the frame 8 to the frame 13 of the given video sequence*

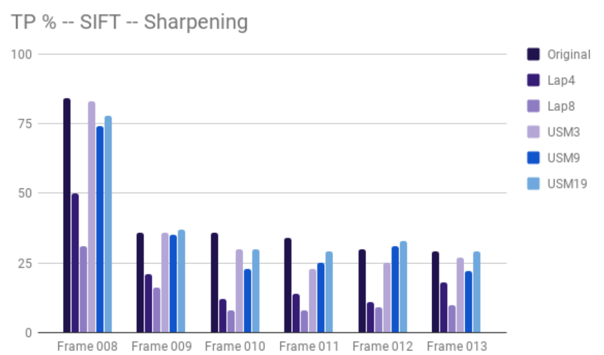

Figure 12. *True positive percentages of SIFT given by frame and by sharpening technique used. The sub-image used has been extracted from the frame 8 and the tests have been applied from the frame 8 to the frame 13 of the given video sequence*

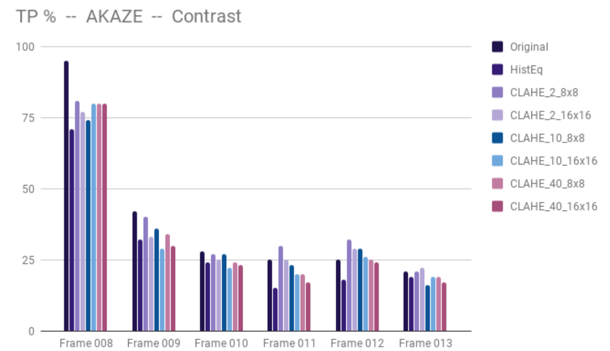

Figure 13. *True positive percentages of AKAZE given by frame and by contrast enhancement technique used. The sub-image used has been extracted from the frame 8 and the tests have been applied from the frame 8 to the frame 13 of the given video sequence*

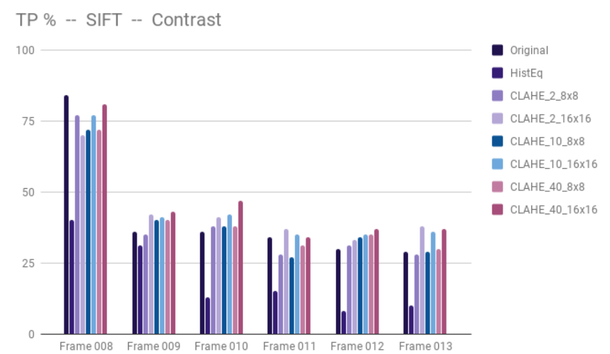

Figure 14. *True positive percentages of SIFT given by frame and by contrast enhancement technique used. The sub-image used has been extracted from the frame 8 and the tests have been applied from the frame 8 to the frame 13 of the given video sequence*

AKAZE test, the TP percentages are almost always lower than the ground truth. However, the number of inliers obtained for contrast enhancement using CLAHE in the SIFT test, tends to increase from the *frame n* to *n+5*. Though subjective, it can be observed that, depending on the feature detector used, worse image visually can correlate with better the TP rate. This is an interesting result, as it indicates a perhaps direct conflict between what one might call a high quality image for human viewing and computer vision performance.

Pedestrian Detection

When presenting the pedestrian detection algorithm, we elect to differ slightly from the feature extraction. We present it more from a design perspective of searching for the best ISP for pedestrian detection by maximizing the overall performance through configuration. Hence we present below some "best configurations".

Before applying sharpening and contrast enhancement, the pedestrian algorithm has been run for the entire catalog without any image enhancement pre-processing, so the TP and FP values obtained are considered as ground truth (original values). Then, both sharpening and contrast have been applied prior to the pedestrian detection algorithm. Just to limit the configuration space, we pick one of the sharpening and contrast enhancement techniques. For sharpening, it has been chosen to use the Laplacian technique. This filter has only one parameter that has two possibilities (*Lap4* or *Lap8*). For contrast enhancement, the CLAHE filter is applied. This filter has two parameters: clipLimit: $[1,15]$ and tileSize: 8×8 or 16×16. Thus, here was 60 possible configurations, all of which have been tested. The objective here is to optimize the PD KPIs with the aim of maximizing the TP rate while maintaining as low as possible FP rate. To achieve this objective, we propose to calculate a *compromise* value (Γ) in order to jointly optimize TP rate (r_{TP}) and FP rate (r_{FP}) values:

$$\Gamma = r_{FP}(1-r_{FP})$$

This test has been made over a catalog of 93 videos. The graph of the Figure 15, shows these three values per configuration. It can be observe that the TP rate value vary from 0.68 up to 0.83. The FP per frame also varies significantly, from 0.039 to 0.101.

In Table 1, we establish the best configurations by TP rate and compromise value. The best TP found in the configuration is 0.83, which is about 0.045 increase over the original. By looking at the FP per frame value, however, it can be observe that it is

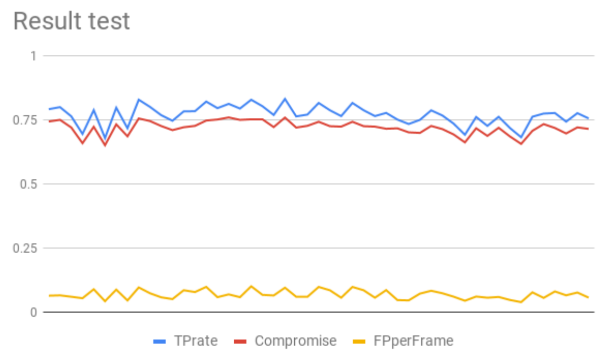

Figure 15. *Chart showing TP, FP and compromise values per configuration. 49 configurations are represented. Blue line: TP rate, Red line: compromise value, Yellow line: FP per frame*

	Original	Best config (by r_{TP})	Best config (by Γ)
Γ	0.7451	0.7589	0.7595
r_{TP}	0.7869	0.83	0.81
r_{FP}	0.055	0.095	0.069

Table 1. *TP rate, FP rate and compromise values for original catalog, best configuration by TP rate and best configuration by compromise value*

almost doubled (0.095) compared to the original. However, the compromise value takes in account the TP rate and the FP per frame, and is perhaps a better measure. The last column shows the KPIs values if sorted by compromise value, the best TP rate is 0.81, which is more that 0.025. In this case, the FP per frame is still low (0.069).

	Best config (by r_{TP})	Best config (by Γ)
Lap	Lap4	Lap8
clipLimit	2	2
tileSize	8x8	8x8

Table 2. *Parameters of the best configurations*

Table 2 shows the parameters value of the best configuration found when the sort has been made per TP rate or per compromise value. Figure 16 shows an image processed by the parameters value given by the best configuration found when the configuration are sorted by compromise value. It can be observed that the "image quality" of this image does not correspond to a good looking image for human viewing applications.

Conclusion

The study presented in this paper has shown that there is a measurable and significant impact on computer vision when processing that models ISP is applied to the test set, albeit acknowledging limitations in terms of the model and data set. However, it is a strong indicator that ISP tuning for computer vision is a significant potential area of investigation in order to obtain the optimal performance of computer vision algorithms, both traditional feature extraction and machine learning algorithms. Interestingly, we have observed that there can be a direct conflict between the tuning of ISP that is optimal for computer vision, and that which is pleasing for human visual consumption. With the continued

Figure 16. *Image sharpened and contrasted using the parameters given by the best configuration found when sorted by compromise value (Lap8, clipLimit=2 and tileSize=8x8)*

push towards performance maximization of CV perception for automated driving the optimization of all parts of the processing pipeline have to be considered thus ISP tuning for CV can simply not continue to be ignored.

Future work will perform the study on raw images captured at image sensor level to allow full control of ISP parameter tuning/testing through ISP simulator or hardware replay. It is planned to broaden the study with other ISP processes, lens abnormalities, use cases and algorithm types and make steps towards automated tuning of ISP specifically for computer vision.

References

[1] 1858-2016 - IEEE Standard for Camera Phone Image Quality. Technical report, IEEE Camera Phone Image Quality Working Group, 2017.

[2] IEEE P2020 Automotive Imaging White Paper. Technical report, IEEE P2020 Working Group, 2018.

[3] M. Heimberger, J. Horgan, C. Hughes, J. McDonald, and S. Yogamani. Computer vision in automated parking systems: Design, implementation and challenges. *Image and Vision Computing*, 2017.

[4] J. Horgan, C. Hughes, J. McDonald, and S. Yogamani. Vision-based driver assistance systems: Survey, taxonomy and advances. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2015.

[5] L. Yahiaoui, J. Horgan, S. Yogamani, C. Hughes, and B. Deegan. Impact analysis and tuning strategies for camera image signal processing parameters in computer vision. In *Irish Machine Vision and Image Processing Conference (IMVIP)*, 2018.

[6] S. A. K. Tareen and Z. Saleem. A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK. In *International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2018.

[7] D G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[8] P. F. Alcantarilla, A. Bartoli, and A. J. Davison. Kaze features. In *European Conference on Computer Vision (ECCV)*, 2012.

[9] P. F. Alcantarilla, J. Nuevo, and A. Bartoli. Fast explicit diffusion for accelerated features in nonlinear scale spaces. In *British Machine*

Vision Conference (BMVC), 2013.

[10] W. McIlhagga. Estimates of edge detection filters in human vision. *Vision Research*, 2018.

[11] P. G. J. Barten. *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. Society of Photo-Optical Instrumentation Engineers (SPIE), 1999.

[12] A. M. Reza. Realization of The Contrast Limited Adaptive Histogram Equalization (CLAHE) For RealTime Image Enhancement. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, 2004.

[13] J. Canny. A computational approach to edge detection. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 1986.

Author Biography

Lucie Yahiaoui received her M.Sc. in Image and Signal processing from the University ESEO, France (2017). Since then she has worked as a Computer Vision Research Engineer in the Research and Development department at Valeo Vision Systems (Ireland).

Jonathan Horgan is a Computer Vision & Deep Learning Architecture Manager and Expert at Valeo Vision Systems. He has worked in the field of computer vision for over 14 years with a focus over the last 8 years on automotive computer vision for Advanced Driver Assistance Systems (ADAS) and automated parking & driving. He has technically led computer vision projects for many large automotive OEMs on state of the art series production developments. He is now working on next generation advanced computer vision and deep learning with the ultimate goal of achieving fully autonomous driving and parking. He has 17 publications in peer reviewed conferences & journals and over 70 patents filed in the field of computer vision for ADAS and Autonomous driving.

Ciarán Hughes is a computer vision architect at Valeo, focusing on algorithms for surround-view cameras. He has 15 years of experience in automotive computer vision and camera design. He has a Ph.D. in Electronics from NUI Galway.

Brian Deegan is a Senior Research Engineer and Senior Expert at Valeo Vision Systems. He received a Bachelor of Engineering in Computer Engineering in 2004, and a Masters in Science in Biomedical Engineering in 2005. He received a PhD in Biomedical Engineering in 2011. His main research interests are in automotive image quality for both viewing and machine vision applications.

Patrick Denny is a Senior Research Engineer and Senior Expert at Valeo Vision Systems in Tuam, Ireland, where he has been a Technical Lead and internal consultant for radiofrequency engineering, camera and imaging system development and production for the last 17 years for the leading automotive marques. He has been Technical Lead for the development of image quality on automotive vision systems for BMW, Landrover, Ford, VW, Daimler and Audi. He is Adjunct Professor of Automotive Electronics in NUI Galway, Ireland. He received his B.Sc. in Experimental Physics and Mathematics from NUI Maynooth in 1993, an M.Sc. in Pure Mathematics from NUI Galway in 1994 and a Ph.D. in Physics in 2000 from NUI Galway and a Professional Diploma in Data Analytics from University College Dublin, Ireland, in 2016. He has over 112 patent applications in 56 patent families and over 30 publications in refereed journals, conference proceedings and books on topics ranging from image quality to planetary physics to artificial intelligence. He co-founded, chairs and is on the steering committee of the AutoSens and Electronic Imaging's Autonomous Vehicles and Machines conferences and co-founded and chaired the IS Auto conference. He has developed and encouraged an international network of technologists in industry and academia in imaging is also a Co-founder and Board Member of the IEEE P2020 Automotive Imaging Standards body.

Senthil Yogamani is a computer vision architect and technical leader at Valeo Vision systems. He is currently focused on research and design of the overall computer vision algorithm architecture for surround-view camera visual perception in autonomous driving systems. He has over 13 years of experience in computer vision and machine learning including 10 years of experience in industrial automotive systems. He is an author of 50 peer reviewed publications and 33 patents. He serves in the editorial board of various leading IEEE automotive conferences including ITSC and ICVES and advisory board of various industry consortia including Khronos, Cognitive Vehicles and IS Auto. He is a recipient of best associate editor award at ITSC 2015 and best paper award at ITST 2012.