PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Enhancing learning-based computer vision algorithms accuracy in sUAS using navigation wide-angle cameras

Xavier Dallaire, Julie Buquet, Patrice Roulet, Jocelyn Parent, Pierre Konen, et al.

Xavier Dallaire, Julie Buquet, Patrice Roulet, Jocelyn Parent, Pierre Konen, Jean-François Lalonde, Simon Thibault, "Enhancing learning-based computer vision algorithms accuracy in sUAS using navigation wide-angle cameras," Proc. SPIE 11870, Artificial Intelligence and Machine Learning in Defense Applications III, 1187009 (12 September 2021); doi: 10.1117/12.2600197

SPIE.

Event: SPIE Security + Defence, 2021, Online Only

Enhancing learning-based computer vision algorithms accuracy in sUAS using navigation wide-angle cameras

Xavier Dallaire[1], Julie Buquet[1,2], Jocelyn Parent[1], Pierre Konen[1], Jean-François Lalonde[2], Simon Thibault[1,2]

[1]Immervision Inc, 2020 Robert-Bourassa Blvd., Suite 2320, Montreal, Quebec, H3A 2A5, Canada
[2]Centre d'optique, photonique et laser, Département de physique, de génie physique et d'optique, Université Laval, 2375, rue de la Terrasse, Quebec, G1V 0A6, Canada

ABSTRACT

The new generation of sUAS (small Unmanned Aircraft Systems) aims to extend the range of scenarios in which sense-and-avoid functionality and autonomous operation can be used. Relying on navigation cameras, having a wide field of view can increase the coverage of the drone surroundings, allowing ideal fly path, optimal dynamic route planning and full situational awareness. The first part of this paper will discuss the trade-off space for camera hardware solution to improve vision performance. Severe constraints on size and weight, a situation common to all sUAS components, compete with low-light capabilities and pixel resolution. The second part will explore the benefits and impacts of specific wide-angle lens designs and of wide-angle images rectification (dewarping) on deep-learning methods. We show that distortion can be used to bring more information from the scene and how this extra information can increase the accuracy of learning-based computer vision algorithm. Finally, we present a study that aims at estimating the link between optical design criteria degradation (MTF) and neural network accuracy in the context of wide-angle lens, showing that higher MTF is not always linked to better results, thus helping to set better design targets for navigation lenses.

Keywords: Wide-angle, sUAS, machine vision, AI, machine learning, machine perception

1. INTRODUCTION

Recently, UAVs and others autonomous vehicles have democratized. To make such device able to navigate in our world, it must be aware of its 3D environment and adapt its trajectory according to it. Severe constraints on size and weight, a situation common to all sUAS components, compete with low-light capabilities, high pixel resolution and high surroundings coverage, which are desirable for increased navigation capabilities. These constraints and trade-offs led to a variety of solutions currently available on the market that will be discussed in the next section.

Tasks related to navigation are not performed by human, but via computer vision algorithm. Thus, is it important to understand how effects that are typically linked with image degradation like lower MTF, or distortion truly affect aspects of the computer vision pipeline. Using wide-angle images, especially relevant because of their ability to capture a great amount of the device's surroundings, we determine the effect of different optical characteristics on computer vision, specifically for neural networks for object identification. In the third section, we study the influence of distortion rectification on pre-trained neural networks.

In the fourth section, we present a study that aims at estimating the link between optical design criteria degradation (MTF) and neural network accuracy. By exploring the effects of image degradation on the computer vision pipeline, we can better identify specifications for optical designers allowing to specify optical design requirements on typical metrics (MTF, PSF) to obtain the desired performances for the computer vision task considered.

2. HARDWARE TRADE-OFFS FOR LOW-LIGHT WIDE-ANGLE SUAS NAVIGATION CAMERAS

The sUAS space is extremely fragmented in term of navigation sensing technologies available. This points out the fact that there is no one does it all solution and that trade-offs are necessary. We can identify three main competing requirements which are ease of integration (size, weight, power consumption), coverage (FoV), and sensing accuracy (range, precision, operating scenarios).

Artificial Intelligence and Machine Learning in Defense Applications III, edited by Judith Dijk, Proc. of SPIE Vol. 11870, 1187009 · © 2021 SPIE
CCC code: 0277-786X/21/$21 · doi: 10.1117/12.2600197

The principal challenge for sUAS is to comply with the operational requirements, the most important of all being to maintain a small size and weight. Each addition in weight will impact power consumption and affect navigation. Each addition in size will result in a bigger, less maneuverable drone.

When cameras for drones are discussed, main payload cameras on gimbals are often referred at. They enjoy a privileged place with a considerable allocation of the space and weight dedicated to them in the general structure. This is usually not the case for navigation camera, at least not yet. Navigation cameras are currently expected to be extremely small (≤ 20 mm height for full camera module) and light (≤ 5 g for full camera module) so that the integration on the main frame of the drone can be possible. Because the navigation cameras are meant to provide a wide coverage around the aircraft as well as 3D information, multiple cameras (4 ~ 8) are necessary, increasing even further the burden of keeping a small size and weight.

The immediate result of this size limitation is a drastic reduction in available sensor technology. High performance low-light sensors are typically designed with bigger pixel, which mean that to obtain a higher resolution, the sensor, and thus the lens paired with it, tend to have a size too big to be a realistic prospect for being a sUAS navigation camera. Solutions that are deployed currently for low-light operation are varied and imperfect, as discussed below. They reflect the fact that the market is divided on performance target for sUAS. The solutions are presented here as a high-level summary of what is on the market.

The most widespread solution for enabling performance in low-light sense-and-avoid capabilities is onboard illumination. Either being visible light, near-infrared, or a mix of both, light emitter is integrated on the drones enabling cameras with limited low-light performances to function in darker environments. This solution brings considerable downsides because of the additional power requirement and weight.

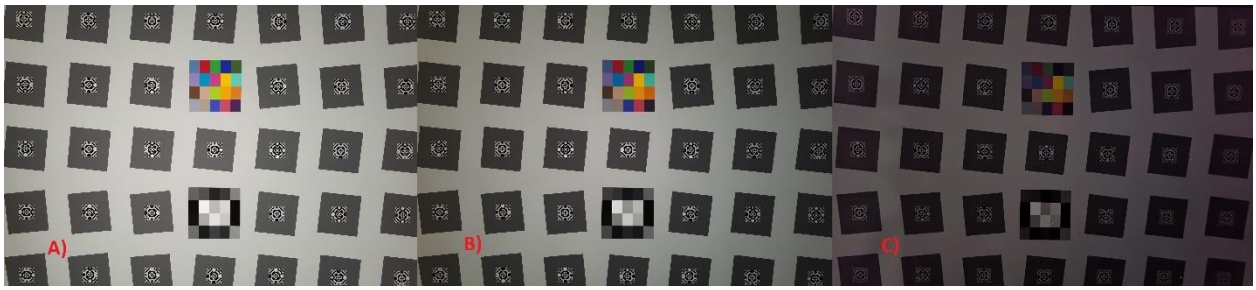
Another solution observed is to forfeit coverage. In certain cases, sense-and-avoid functionality in low-light will be limited to fewer instruments, for example a thermal camera on the payload, that can only cover a small portion of the drone's surroundings. The effect is to provide limited autonomy and to increase the reliance on a human pilot.

GPS navigation is often used to facilitate the drone autonomy. It is well-suited to night operation but is very limited outside of a sky clear of obstacles. GPS cannot provide the relevant information to safely take-off and land, even less fly reliably inside structures or in undergrounds tunnels.

Scanning LIDAR are sometimes used as part of the main payload. They can map the 3D environment and due to their scanning nature, can cover much of the surroundings of the drone. However, today, the important weight and size that these systems usually have makes them an unideal choice for sense-and-avoid functionality on drones.

Time-of-flight, also called flash LIDAR is also used. Their small size makes them attractive for integration. Since the light is sent in the whole FoV at once, the power requirement is increased or, as we usually observe, range reduced significantly.

Another versatile option which we are going to explore in more detail here is the standard miniature wide-angle camera as it presents interesting characteristics in term of size and coverage [1]. A typical downside for higher resolution miniature sensors is a certain drop in low-light performances compared to bigger sensor with wider pixels. A careful balance must be achieved between resolution and pixel light-gathering capabilities. Taking a wide-angle smartphone lens as a general example in Figure 1 below, we can observe that when operating in well-lit condition, the system comprising of the lens, the sensor and ISP will produce a mostly constant MTF output. However, reduced lighting will result in decreased contrast and added noise to the image, thus lowering the MTF. For the purpose of this paper, these effects will be represented in first approximation as MTF degradation only and discussed in Section 4.



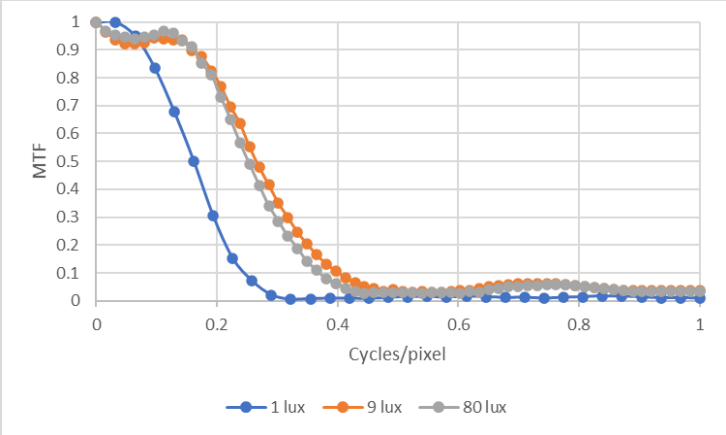


Figure 1. On top, a smartphone using the IMX608 sensor for A) 80 lux, B) 9 lux and C) 1 lux. Below, MTF is shown for the top right corner of each image. The corner is the part of the image most susceptible to impact caused by low illumination du to its reduced relative illumination in typical miniature wide-angle lens.

Miniature wide-angle lenses can offer small size and weight while also providing high coverage and performances for sense-and-avoid functionality in sUAS. Low-light performances can be increased through careful selection of the sensor (lower resolution and better light-gathering capabilities), small f/#, increased spectral band coverage and controlled relative illumination. A set of camera module following this design philosophy is in production right now and will provide the basis for further experimental studies.

Covering the visible spectrum and potentially the NiR, the wide field of view paired with the distortion found in this kind of lens also present additional challenges and opportunities for machine vision which we will explore in the next section. When defining the required specifications for this lens type, typical benchmarks are based on human vision. In the last section, we will explore how performance metric can directly impact the computer vision task considered and thus influence design targets.

3. INFLUENCE OF DEWARPING ON NEURAL NETWORK ACCURACY

In our first experiment, we aimed at understanding and quantifying distortion rectification impact on learning-based applications. In our case study of object identification, deep convolutional neural networks are trained to detect and label objects from a single input image. To do so, networks extract useful information by processing a succession of convolutional and non-linear calculations. This method has proven to be very efficient in multi-scale feature extraction for various tasks such as object detection, depth estimation, face recognition and 3D reconstruction.

A convolutional kernel of constant size is applied on the entire image at each layer of the network. In this context, using CNNs implies a translational invariance on the image studied. More precisely, it requires the magnification to be constant along image direction. However, this assumption is only true for images represented by the pinhole camera model. Such images have limited field of view and will be referred as perspective images.

On the contrary, traditional wide-angle images have a magnification that will vary from the center toward the edges of the image leading to decreasing accuracy when used on neural networks, especially the ones pre-trained on perspective images. Indeed, the strong apparent distortion will bend lines and modify object proportion. This effect increases with the field of view.

However, such wide-angle cameras are highly interesting for computer vision community because of their ability to capture the entire surroundings of a sUAS with a minimal number of camera modules reducing both weigh, power consumption and cost. To be able to analyze such images while keeping acceptable performances, a lot of work have been devoted to estimating and rectifying the distortion. With the camera intrinsic parameters, it is possible to recover the 3D coordinates in the real world of a pixel on the image. By estimating the distortion parameters, it is then possible to reproject this 3D point by considering such rectification resulting in a distortion free perspective image. Such process however leads to pixel stretching on the edges of the image and a crop in the field of view is often needed.

Additionally, other popular reprojections are investigated for wide-angle images like LatLong panorama and the Mercator projection because they represent the entire field of view. Even if this kind of projection cannot fully rectify the distortion, the apparent proportion are pleasant for human perception [2]. For this reason, it has become popular to develop new and more complex panoramic projection models which aim at reducing the apparent distortion [3,4,5].

With the democratization of automatic vision task such as object identification, we can wonder if this kind of reprojection, which helps to improve human perception using wide-angle images, has the same effect on neural network accuracy. Indeed, the translational invariance mentioned before is still not respected but the magnification variations are drastically reduced which should help to get closer to performances obtained on distortion free images.

In our study, we evaluate the improvement induced by such reprojection on object detection. More precisely, we compared the accuracy obtained by Yolov4 [6] pre-trained neural network on both datasets: unrectified and rectified. Both datasets containing around 1500 images. The first one contains unrectified wide-angle images captured in the streets of Montreal from a car. The camera used was a panomorph lens with a field of view of 180°. For the second dataset, we used a distortion rectification algorithm on the first dataset to obtain images on perimeter projection. We tested our algorithm on our private dataset for qualitative results and PixSet dataset for quantitative study [7]. Both were composed of pictures taken with panomorph lenses from the front of a car.

We used yolov4 pretrained on narrow-angle dataset MsCOCO [8]. This dataset consists in 328 000 narrow-angle images with 80 different classes of objects labelled. We chose to use yolov4 because it is widely used in different industries and its architecture have proven very efficient. The pre-trained option was chosen to maximize that amount of data available for this study.

We sampled our datasets and took around 1200 images to realize the experiment. For each image and each object detected, yolov4 outputs its label, the coordinates of the bounding box, and the confidence score associated with the detection. We compared each detection with the ground truth annotation and measured the performances using precision and recall.

Each object detection was categorized either as true positive (TP) if the object is detected by Yolov4 and labelled similarly to the reference annotation, or false positive (FP) if it's detected but not correctly labelled or even not present in the reference objects detected. Each detection was weighted by the confidence score output, like that a correct detection with a higher score will increase more importantly the network accuracy. In the same way, a false positive with a higher score constitute a bigger mistake from the network and will then have a bigger impact on the accuracy drop. Then, we counted the missed detection as false negative (FN). From the total number of TP, FP and FN representing all detections, we could compute two different metrics to evaluate yolov4 accuracy. They are summed up in Figure 2 below.

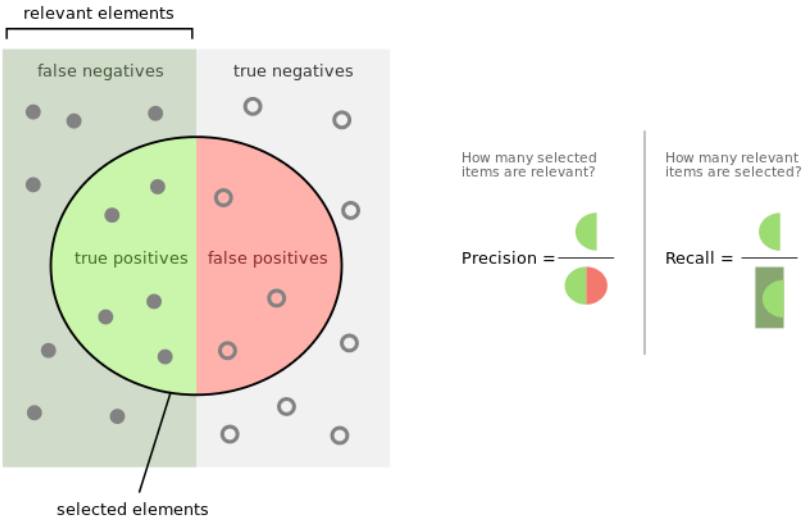


Figure 2. Visualization of Precision and Recall

First, precision represents the ratio between the number of objects correctly detected (TP) and all the object detected (TP+FP). This metrics will then evaluate the number of hallucinated objects by yolov4 compared to the total number of detections performed by the network

On the other hand, measuring the recall allow us to quantify the sensitivity of the network by computing the ratio between the number of objects correctly detected (TP) and the total number of objects to detect (TP+FN). This metric will highlight the number of missed detections compared to all objects detectable, according to the reference annotations, on the images. We did not count non-detected objects that were occluded as they were hidden on the image. We computed both metrics for both datasets and obtained results summarized in Table 1.

Table 1. Results obtain with MsCOCO narrow-angle pretrained Yolov4 on PixSet for object identification

	Non-rectified	**Rectified**	**Relative Improvement**
Precision (%)	70.5	74	+ 5.0%
Recall (%)	35	37	+ 5.7%

For both datasets, we can notice that the overall performances on both metrics are very low compared to state-of-the-art results. This can be explained by the structural difference between images taken for training and our test datasets. We still obtained results comparable to other similar studies on wide angle images [9]. Moreover, we can see that the results on the rectified dataset are better than in both precision and recall. We found out that the relative improvement for both metrics reached at least 5% which is non-negligeable, especially in application such as autonomous driving where a missed detection can be a pedestrian crossing the street.

We completed our study with a qualitative analysis of the results. By analyzing the images, we found out that most missed detection on non-rectified images that are correctly detected on the rectified dataset happen on the edges of the image where the distortion is the most noticeable. Moreover, it was harder in the first dataset to detect further or truncated objects or those under challenging lightning conditions as shown in Figure 3 below. We will further investigate in this direction with a quantitative spatial analysis of the performances.

After this study, we concluded that distortion rectification can help improving neural network accuracy. Because of this process, the images have a structure closer to those used in the training dataset. However, this process remains time consuming, and the perimeter view is still limited to a 140 degrees field of view. With the democratization of wide-angle imaging systems, it is getting easier to collect dataset of such images and some works have tried to directly use them for training. For example, PixSet dataset provides wide-angle images annotated for different tasks from 3D point cloud estimation to object classification.

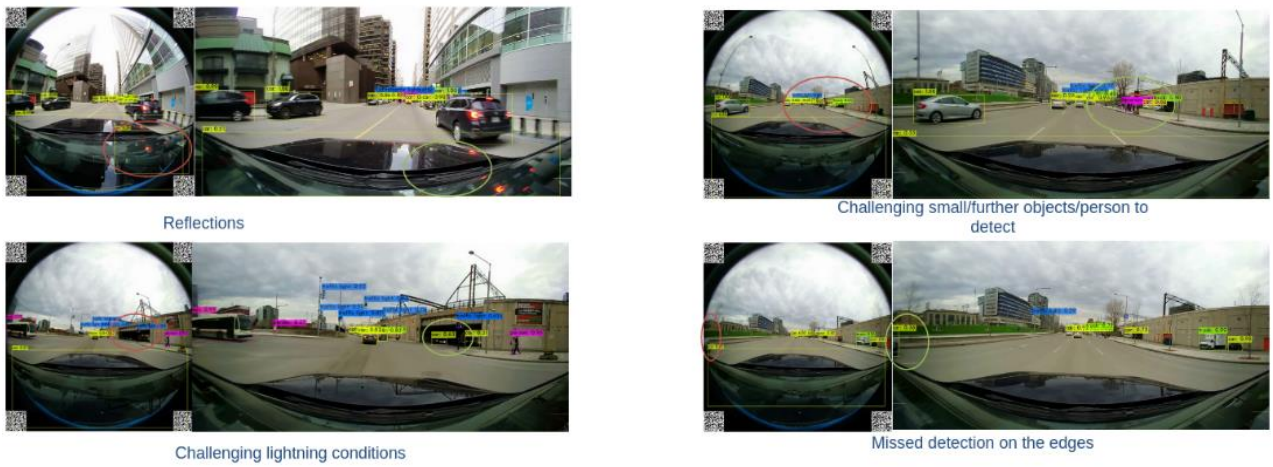


Figure 3. Examples of challenging detection on non-rectified images from our dataset

With such advance, recent works [8] focus on directly training neural networks on wide-angle images to keep the entire field of view analyzed without any post processing operation. Even if performances remain lower than for perspective images, some works showed interesting results on such images. Particularly, some work [10] focus on evaluating the

impact of nonlinear distortion for single image depth estimation. They simulated images from three wide-angle system with different distortion functions. Such lenses can capture a field of view of 180 degrees, as a fisheye presents a linear distortion, panomorph lenses introduce non linearities to induce augmented resolution on a region of interest (ROI) of the image. By training identical networks for single image depth estimation, with images simulated from these different systems, they showed that nonlinear distortion can help improving neural network accuracy. By using panomorph lenses with a specific region of interest during the training, the network accuracy was increased within this ROI. Thanks to that, it is possible to locally improved performances.

In the same direction, other recent works focus on linking optical design characteristics and metrics to computer vision tasks [11]. Indeed, metrics such as PSF or MTF are used by optical designers to evaluate the image quality provided by an optical system. So far, such metrics are optimized to fit a pleasant image for human perception, but it is still not clear how does a neural network will precisely react to an MTF degradation for example. In a recent experiment, we tried to understand and quantify the impact of PSF degradation on pretrained yolov4 for object detection as before. The goal is to help optical designer to determine characteristics for their design that will ensure the required performances for a computer vision application.

4. EFFECT OF IMAGE DEGRADATION OF MACHINE VISION ALGORITHM PERFORMANCES

To realize such experiment, we simulated images having different PSF from the same scene. We used PixSet to get wide-angle images. We considered that the images were perfect, meaning that the PSF was diffraction limited. We simulated the degradation due to MTF modifications using the PSF as described below.

We can entirely simulate the rendering through an optical design using the PSF. Indeed, a PSF represents the image corresponding to a point in the object space. It will give us information on the shape, and the lightning on the camera sensor. It is then possible to simulate the effect of an optical design by convolving the initial image, representing the scene to image, with this PSF. In this 2D convolution, the PSF will act like the convolutional kernel which size in pixel will be adjusted to fit the size of a pixel on the sensor. Each pixel of the initial image will be convolved to simulate the corresponding airy disk on the sensor. However, a PSF is spatially varying. Off-axis aberration for example, will appear only for wider field of view. To process the convolution, each pixel must be convolved with the PSF corresponding to its field of view. In this way, for a single image, different PSF must be estimated and a sampling in the field of view must be done.

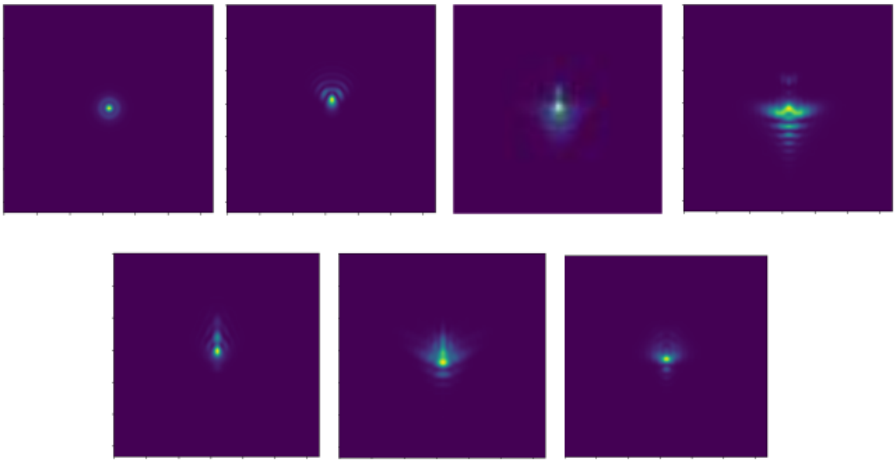


Figure 4. Representation of the PSF used to simulate MTF degradations (MTF degrades from top to bottom, left to right)

Our work aims at estimating the impact of PSF degradation on object detection. The first approach we chose was to apply the same PSF on the entire field of view. We were then able to have preliminary results even if the degradation will be uniform on the image.

For our experiment, we generated 7 different PSF, shown in Figure 4, from an optical design. To do so we simply estimated the PSF at different field of view for this design. Each PSF was represented by an 128x128 RGB image. We had access to the width in micron for each of them. Before applying such PSF on the image, we had to resample them to match the resolution of a sensor. We generated five different resolutions: 14x14, 30x30, 45x45, 60x60 and 75x75. 14x14 was chosen to fit a pixel size on the sensor of 3,5um used in PixSet set up. Then we simply increased the resolution to see the effect of a wider PSF, meaning a bigger degradation.

In this way, we could generate 35 different datasets composed of 200 images, 5 for each of the seven PSF by doing a 2D convolution on the entire image. We generated images in gray scale from PixSet dataset as shown in Figure 5.

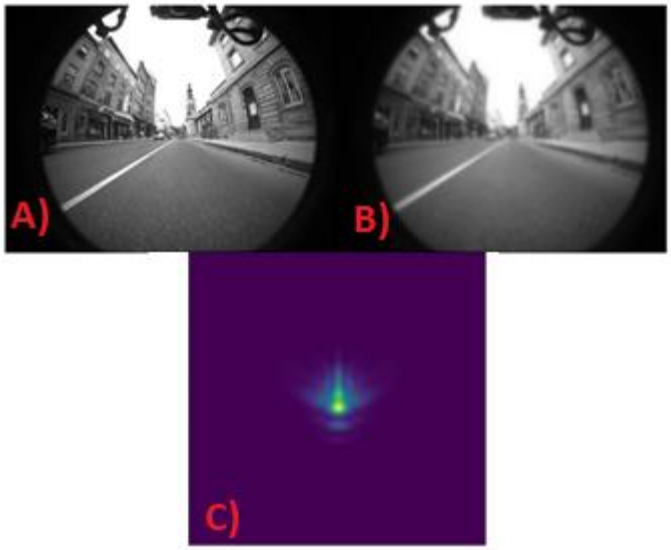


Figure 5. A) Original PixSet image. B) Example of a degraded image. C) PSF used for uniform convolution with A) to obtain B) with a 128x128 resolution.

The top left PSF represented in Figure 4 corresponds to a diffraction limited system, it will be considered as the ideal PSF. We took the value of the MTF for each of the generated PSF and could compute the relative MTF degradation for each PSF taken. We tested yolov4 on each dataset and used the same metrics as before, precision and recall evaluating the network accuracy. The F1 score corresponds to an average of both metrics:

$$F1 = \frac{2PR}{P+R} \quad (1)$$

When we look at the results presented in Figure 6 below, we can first notice that for a single PSF, the performances are lower when the resolution taken is higher (as a higher resolution means a bigger PSF) but they all follow the same trend. For further studies we then just consider one of the resolutions for each PSF. After computing the results, we calculated the relative drop in network accuracy compared to the reference PSF as a function of the relative MTF degradation. We found out that the network accuracy decreases with MTF degradation. Particularly, F1 score has a relative drop of 3% (relative F1 score) when MTF degradation is around 17% and reaches 4.5% when the degradation reaches 40%.

Figure 6. Quantitative results. From top to bottom: Precisions, Recall and F1-score obtained on a subset of PixSet. For each row, the absolute values are presented on the left and the relative drop on the right

An interesting thing to notice in this study is that for the precision, the MTF degradation leads to better results even if the improvement was not significant compared to the other metrics (below 1%). This can be explained by the fact that, with blurrier image, yolov4 will not notice further objects that can be detected but misclassified with a better MTF. This should be confirmed by testing more images in the future.

With this study, we showed that a better MTF tends to improve neural networks accuracy even if it can lead to additional misclassification. It is interesting to notice that by computing the results as a function of the MTF, we only consider the contrast on the image. However, we could notice than even with a low MTF degradation, a PSF with higher geometrical degradation led to a significant drop in the performances. This is noticeable for the PSF leading to 26% of MTF degradation. We can see that the performances are lower than for the rest of the PSF. However, if we look at the corresponding PSF in Figure 4, we can see that it contains much more aberrations than the others. The PSF is wider and more distorted even if the degradation in the MTF is reasonable.

This last observation let us think that the MTF has a lower impact than aberrations in learning-based application which confirms the results seen in [11]. They also observed that for this kind of application, the performances are optimized with a sharper PSF on the region of interest for example on the top of the image and on the side in the case of traffic light identification. It would be interesting to investigate this observation in the future by improving the image simulation. By computing a spatially varying distortion we will be able to observe the relative performances on the same image at different field of view.

5. CONCLUSION

Miniature wide-angle lenses are more commonly used for autonomous vehicle navigation as they can offer small size and weight while also providing high coverage and performances for sense-and-avoid functionality in sUAS. Design targets like distortion, MTF, PSF can be controlled at the design level to maximize machine vision algorithm performances, making it the optimal solution. Additionally, low-light performances can be increased through careful selection of the sensor (lower resolution and better light-gathering capabilities), small f/#, increased spectral band coverage and controlled relative illumination. A set of camera module following this design philosophy is in production right now and will provide the basis for further experimental studies.

Distortion and MTF degradation impact on learning-based computer vision algorithm were explored. It was shown that distortion can lead to reduced accuracy in the case of neural network trained for narrow-angle images, but that in the case of especially variant distortion across the field, training new networks can lead to increased performances in specific part of the field of view compared to dewarped images. Effect of MTF and PSF were explored, showing a significant dependency on PSF shape, not always visible from a MTF representation standpoint. Future work will cover spatially variable PSF on the same image, thus allowing a better representation of an actual optical system and guiding optical design parameters definition for machine vision systems.

REFERENCES

[1] Simon Thibault, Jocelyn Parent, Hu Zhang, Xiaojun Du, Patrice Roulet, "Consumer electronic optics: how small can a lens be: the case of panomorph lenses," Proc. SPIE 9192, Current Developments in Lens Design and Optical Engineering XV, 91920H, 2014

[2] 3D Object Detection from a Single Fisheye Image Without a Single Fisheye Training Image, Elad Plaut, Erez Ben Yaacov, Bat El Shlomo, In Workshop on Omnidirectional Computer Vision (CVPRW), 2021

[3] Practical Wide-Angle Portraits Correction with Deep Structured ModelsJing Tan, Shan Zhao, Pengfei Xiong, Jiangyu Liu, Haoqiang Fan, Shuaicheng Liu, arXiv:2104.12464CVPR 2021

[4] YiChang Shih, Wei-Sheng Lai, and Chia-Kai Liang.Distortion-free wide-angle portraits on camera phones.ACMTrans. Graphics, 38(4):1–12, 2019

[5] J. Parent and S. Thibault, "Controlled distortion," *2012 11th Euro-American Workshop on Information Optics*, 2012, pp. 1-3, doi: 10.1109/WIO.2012.6488921.

[6] "YOLOv4: Optimal Speed and Accuracy of Object Detection", A. Bochkovskiy, C. Wang, H-Y. M. Liao, 2020 arXiv:2004.10934

[7] PixSet: An Opportunity for 3D Computer Vision to Go Beyond Point CloudsWith a Full-Waveform LiDAR Dataset, Jean-Luc Deziel, Pierre Merriaux, Francis Tremblay, Dave Lessard,Dominique Plourde, Julien Stanguennec, Pierre Goulet, and Pierre Olivier, White paper LeddarTech, 2021

[8] "Microsoft COCO: common object in context", T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. Lawrence Zitnick, P. Dollár, CVPR 2015

[9] "Automatic Distortion Rectification of Wide-Angle Images Using Outlier Refinement for Streamlining Vision Tasks", V. Kakani, H. Kim, J. Lee, C. Ryu and M. Kumbham, Tasks. Sensors. 2020;20:894. doi: 10.3390/s20030894

[10] Evaluating the Impact of Wide-Angle Lens Distortion on Learning-based Depth Estimation Julie Buquet, Jinsong Zhang, Patrice Roulet, Simon Thibault, and Jean-François Lalonde In Workshop on Omnidirectional Computer Vision (CVPRW), 2021

[11] Differentiable Compound Optics and Processing Pipeline Optimization for End-to-end Camera Design Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-François Lalonde, and Felix Heide ACM Transactions on Graphics (to be presented at SIGGRAPH), 2021