2020 NDIA GROUND VEHICLE SYSTEMS ENGINEERING AND TECHNOLOGY SYMPOSIUM
AUTONOMY, ARTIFICIAL INTELLIGENCE, AND ROBOTICS TECHNICAL SESSION
AUGUST 11-13, 2020 - NOVI, MICHIGAN

BALANCING BETWEEN COMPUTER AND MACHINE VISION - A DESCRIPTION OF AN IMAGERY TOOLCHAIN FOR COMPLEX SOLUTIONS

Paul Skentzos[1], Stephen Pizzo[2]

[1]CTO, Consolidated Resource Imaging, Grand Rapids, MI
[2]Imaging Scientist, Consolidated Resource Imaging, Grand Rapids, MI

ABSTRACT

Machine learning (ML), artificial intelligence (AI), and computational photography (CP) are pushing the boundaries of how electro-optical (EO) and infra-red (IR) sensors are being used. Especially within military environments, users are asking much more from EO and IR sensor suites. While hardware capability continues to advance the state of the art, software has become the true differentiator for how well these sensor platforms perform for the warfighter. This paper presents work that Consolidated Resource Imaging (CRI) has been developing in the areas of machine learning and computational photography. In this effort, we will discuss two areas of understanding: imagery meant for machine vision and imagery meant for human consumption. We will show how the intersection of machine learning and computational photography allow the symbiotic relationship between the human and the computer.

Citation: A. Paul Skentzos, B. Stephen Pizzo, "Balancing Between Computer and Machine Vision – A Description of an Imagery Toolchain for Complex Solutions", In *Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium* (GVSETS), NDIA, Novi, MI, Aug. 11-13, 2020.

1. INTRODUCTION

Computational imaging tools have revolutionized digital imaging, allowing information to be captured and details to be resolved that once seemed only possible in science fiction. As consumer digital cameras and smartphones evolve with smaller more powerful electronics and increase in computational power, computational tools have become an integral part of modern cameras. Algorithms are used to reduce noise, extend dynamic range, improve skin tone rendition and general color fidelity. In these examples the computational tools are part of the system and are executed automatically and invisibly as images are captured. In the scientific community, computational tools are also a part of the imaging system however they are typically executed after the fact in a post processes performed minutes, days or even months after capture. The post processes are normally agnostic with respect to camera and optics selection and even capture settings. The algorithms used in the

different scientific disciplines are often quite sophisticated compared to their consumer versions and that sophistication requires significantly more processing power than has not been possible in consumer/mobile devices until very recently. The multiple high-resolution uncompressed images in the data sets require a lot of memory and the execution speed of the sophisticated algorithms is directly proportional to the processing speed and parallel computing capacity. Another reason the scientific computational tools are executed after the fact is the necessity of an operator to set the multitude of parameters, testing the data sets to find the right balance between variables. Among the scientific disciplines such as astronomy and microscopy, the computational processes exist as a wholly separate function from the image acquisition where an abundance of computing power is available. The extra horsepower allows even more complex computational analysis that can aid in identifying the point spread function (PSF) of the blur/distortion. Once this is known the correct algorithms and settings are easy to identify. Applying this capability in real or near real time for use in "buttoned down" vehicles or for rapid analytics requires additional processing.

This paper seeks to propose a third method somewhere between the brute force of trial and error, and the computationally expensive analysis of the captured images. It lays out the architecture for a comprehensive, integrated, end to end system where the sensor, optics, control hardware, storage and processing pipeline all work together as an entity. A total system where acquisition parameters and processing parameters are not just designed to be complimentary, but one informs the other. . . they are mutually dependent.

The first step is to identify the processing objectives, the various types of distortion that we're trying to correct or filter out. These include extreme Low-Light situations where pixel values from multiple underexposed frames are essentially summed build up a single well exposed, sharp image. General noise reduction and bit depth increase will make the output image robust enough to withstand very aggressive deblur processing. The system also accommodates atmospheric blurring from particulate and vapor compression as well as the variable distortions from unevenly heated air cells along the line of sight. Finally, assuming a dataset of well exposed, reasonably clear images with a minimum sample rate, the system can output super resolution images with extremely high detail recovery and a resolution increase up to 50 percent of the original. Next is a description of the system concepts and criteria for selecting system components beginning with the camera and how it differs from those selected for a traditional long-range remote sensing system. The unique camera control functions are described including one of its most critical functions, an asymmetric sensor read-out trigger. This trigger concept will allow a single high frame rate camera to generate real time UHD video for the operator as well as the high-speed sequential datasets for processing. It will also provide the option of capturing variable exposure sequential frames from HDR processing. The PCIe camera interface is described in the context of the low-level image data output format this interface makes possible as well as the platform and electro-optical system (EOS) metadata that makes the automatic processing possible. Following the process chain, the CS2 image data moves over the PCIe BUS to the GPU for transcoding. Depending on how the frames are tagged they're either converted to a real-time video stream or to a format optimized for temporal sampling multi-frame super resolution. Given the potential for overwhelming the data storage infrastructure, a more efficient approach to archiving is presented. This includes a rethinking of how the original capture frames are viewed in this unique imaging solution.

There are hundreds of techniques and processes available for enhancing photographic images and without identifying which ones are appropriate for a set of images and how they should be applied, they are all useless. Based on previous testing of long-range terrestrial images, certain common features have been identified for a specific target distance, level of magnification, pixel size, environmental conditions, etc. By using the EOS and platform metadata a database table can be generated that identifies the ideal stacking and deconvolution methods for each dataset. This metadata driven approach works for both onboard processing as well as post processing.

Since the system being proposed is intended for a high-speed aerial platform, this document explores the benefits and costs of implementing a forward motion compensation mechanism as well as a short analysis of whether it is actually a requirement. While the platform is always moving there are often various objects in frame moving about in different directions and at different speeds. Since each output image (surveillance asset) is made up of multiple input images, the impact of object motion is explored.

There are dozens of stacking and deconvolution methods publicly available for image enhancement, but the ones presented here have been extensively tested together on long focal length terrestrial images. The specific stacking methodology is based on an open source implementation and is laid out step by step in this document. Next up is an outline of the open source deconvolution steps including the specific algorithms, blur kernel types and parameters that have been used for the testing the concept validation. The source code for both is available in a separate document.

There are two fundamental versions of this system; One where the image data and associated metadata are transcoded and formatted for the respective pipelines so that they use minimal storage space even though the ultimate output assets will exceed the detail and quality of conventional output assets of a similar resolution and frequency. The other version includes the same transcoding and formatting as the former but with an onboard processing framework for the advanced pipeline, so the finished assets are completed as the input images are acquired. For the onboard processing version, multi-camera configuration is also presented.

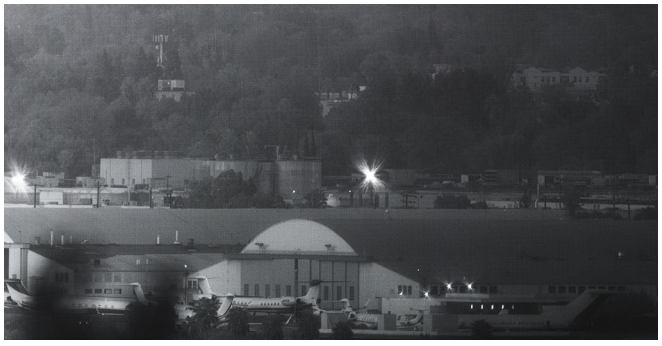
2. Computational Photography Filters

The processing objectives have been broken down into five categories, each requiring specific algorithms and sets of parameters to achieve optimal results. A description of the key ideas are presented.

2.1. Low-light Image Recovery

Normally, capturing an image in low light requires either increasing the gain enough to capture an image at a normal integration or using a long enough integration time to fill the electron well capacity. With the increased gain comes so much noise it’s difficult to differentiate ground truth detail from random noise while the long exposure makes it impossible to capture a sharp image if either the camera or subjects are moving. The processed solution uses a high gain coupled with a very short integration time and multiple quickly captured frames. Fast enough to not only stop motion in a single frame but stop motion across the entire pre-stack sample set. With multiple high-noise, low-light, sharp frames, the light reflected off the ground truth features in the image frame, as low as it is, can be summed from the multiple nearly identical frames in the dataset. The noise however, being rand and so differently distributed in each frame, can be subtracted. The resulting image has the sharpness of a fast integration time with the full rich tonal range

expected in a well exposed image but virtually none of the random noise. Since multiple frames are used, detail can be enhanced as well.



2.2. Noise Reduction and Bit Depth Increase

The second type of correction is Noise-Reduction Bit-Depth Increase. While this is similar to the previous Low-Light process, there is typically enough light for proper illumination however most images still suffer from random noise. While the noise is usually not detrimental to an image for simple viewing, it is absolutely an obstacle to any post sharpening, deblurring, contrast or exposure adjustments, etc. The noise makes an image intolerant to everything but the most subtle adjustments. With the multiple frames that each have full tonal range, the processing is focused on identifying ground truth features and rejecting the random noise. The resulting output image is virtually noise free with well-defined features. While the image will not look significantly sharper after the first phase of processing than any individual input image, it could be described as information rich, with a higher bit depth than the input images so that it can withstand the not insignificant stress of the deblurring/deconvolution processes as well as the regular color and tonality adjustments.

2.3. Atmospheric Distortion

The third type of correction/recovery is Atmospheric Distortion. The atmosphere has a subtle refractive index that changes with temperature. As the sun heats up the ground, rooftops and roadways, the different textures and colors of the objects re-radiate the absorbed thermal energy in localized columns. As these columns rise through the cooler ambient air, they either remain as vertical columns of varied temperature and therefore varied refractive indices or the wind will break them up into varied swirling cells. When the camera system is pointing at an oblique angle, the line of sight must past through the cells and/or columns of unevenly

heated air with vary refractive indices creating a scintillation effect making it impossible to capture any straight lines in the frame and generally creating and image with a random distortion pattern across the frame. Since this distortion pattern is random and constantly changing over time, when multiple high-speed frames are captured, they can be compared to one another to identify the ground truth features then to reassemble in a single more accurate and clear output frame.

2.4. Atmospheric Compression

The fourth type of correction/recovery is Atmospheric Compression. Whether its moisture, smoke, dust or something else suspended in the air, as focal length increases so does the compression of the particulates in the line of site. The effect is to significantly reduce contrast and fidelity as if the camera were shooting through gauze. Even on an apparently clear day, as focal lengths increase, the compression effects become increasingly objectionable. By stacking multiple, quickly captured sequential frames, the common features can be reinforced and built up while eliminating noise and increasing bit dept. This resulting combination can be aggressively deblurred yielding a final output the is very near the diffraction limited ideal.

2.5. Super Resolution

The fifth type of correction discussed here is a version of Super-Resolution where in addition to the typical detail recovery in the other types of correction, it's possible to increase resolution up to 50 percent. To achieve this particular result, certain conditions must be met such as, adequate light for a normal exposure for the integration time and native sensitivity of the camera's sensor. Pushing the sensitivity several stops is fine up to the point where the noise starts to become objectionable in a single input frame. It's also key that there be a high degree of transparency minimal atmospheric compression of suspended particulate. Finally, there needs to be minimal atmospheric distortion from uneven ground radiation. When these conditions are met, all the processing and image reconstruction is starting with what would be described as a set of ideal images with normal noise, adequate exposure and minimal atmospheric degradation. Each frame in the set will contain uniformly resolved details so that the algorithm will have many more consistent segments to use in the stack to build up a cleaner better resolved output. Finally, if the images in the data set are oversampled, it's possible to increase the resolution by as much as 50 percent over the original input images. In other words, using 12MPix input images can yield a high-quality output of 28MPix.

3. Object Detection and Classification

The advent of large format (29 Megapixel and larger) EO machine vision cameras has created an opportunity to develop imaging systems that can cover large areas. The amount of data generated from wide area surveillance in the air and on the ground is truly staggering and each image contains an unbelievable amount of information.

Due to the target-rich nature of overhead imagery and ground based imagery, there is an opportunity to use (ML) algorithms to enhance the value of the imagery. Recently, ML algorithms have begun to be trained for automated target recognition tasks using satellite and overhead imagery. Similarly, this functionality has been demonstrated with ground-based systems. For example, during the last year a competition hosted by ESRI [1] and another by NATO [2] demonstrated how a deep neural network called RetinaNet [3] could be trained to provide excellent automated detection for vehicles. Various auto-tracking techniques have also been applied to EO data demonstrating some ability to track designated vehicles through traffic; however variable lighting and resolution can degrade the process. The development of robust automated target identification algorithms which are capable of tracking targets through variations in resolution caused by conditions varying from morning to late evening, bright and overcast days and changes in altitude greatly enhance the value of overhead imagery and expand mission capabilities.

3.1. Datasets

CRI has petabytes of imagery data captured in the air and on the ground. This same imagery has been captured in the continental US and also outside of the US. The imagery subject ranges from forest fires to college and professional football games to engagements in the Middle East.

CRI is developing this large data set to be used for various ML techniques. The large part of this process is created labeled datasets.

3.2. Transfer Learning

Transfer learning is being used to retrain RetinaNet for various object detection including ground vehicles and marine based objects such as naval vessels. The trained RetinaNet model enables the automated detection of objects in CRI's imagery. Using the labelled dataset, a denoising autoencoder (DAE) is trained to provide image enhancements for low-resolution target images, e.g., to provide a NIIRS5 quality image from a NIIRS4 input image. Once trained, the two ML algorithms will be utilized to process CRI's wide area imagery in two stages. The first processing stage will identify all objects in an image. The second processing stage will provide image enhancement for each vehicle identified by the first processing stage. After both processing stages are complete, a new wide surveillance image will be reconstructed using the image enhanced targets. The process will also include a super resolution convolutional neural network (SR-CNN).

3.3. Image Enhancements

A denoising autoencoder can be trained to provide image enhancement for object targets that are identified in CRI imagery or any imagery for that matter. Denoising autoencoders have been demonstrated to provide image enhancement for noisy images in several industrial applications. We train denoising autoencoders using a large object dataset and can enhancement the imagery resolution. Of large interest, related to this topic, is how to optimize the denoising autoencoder's neural network architecture to provide image enhancement. We have trained a proof of concept

denoising autoencoder with a small subset of data. The denoising autoencoder architecture is being further refined through trial and error testing. Using this mechanism has demonstrate that lower quality target images (e.g., NIIRS4) can be enhanced to provide NIIRS5 image quality.

3.4. Detection Probability

Metrics such as the probability of detection and false alarm rate are evaluated for the RetinaNet ML model. Image quality comparisons are used to demonstrate the denoising autoencoder's image enhancement capability. We have begun the processing requirements of the ML techniques that are being optimized using hardware and software to support real-time processing.

4. Applications

There are a number of applications where the combination of CP and ML can be applied. Large harbor protection could be employed to detect and classify objects at much greater distances. Buttoned up ground vehicles could be navigated by humans or automated with only EO/IR sensors. This technology could also be used to as a sort of "invisible headlight" where a combination of EO and IR sensors are used for navigation in little or no lighting conditions.

4.1. Ground Vehicle Imagery

CRI has developed technology that incorporates *N* cameras in a single field of view. That is, we have developed a system where we can take any number of cameras and stich them into a single field of view image. There is a calibration process has to take into account 29 image regions, 450 measurements, and 176 parameters, making the process computationally intensive. The current calibration process is performed in MATLAB but is now being ported over to the C++ language and making use of graphical processing units (GPUs). The result of the calibration is a configuration file that is read into the system software when starting up.

The original system was intended for stationary, ground based systems, but it is now being developed for use in vehicles. Additionally, more than one system can be used, providing a seamless camera space that can cover an entire outside of a building or can provide for 360 coverage of a moving vehicle.

The image below shows a low-resolution image of the G-iiS system used on a vehicle.

The full image resolution is an 80mp image.

4.2. City Ports and Borders

The ability to bring this advanced CP and ML capabilities to ground based on boarders and ports such as those found around the San Diego area is a logical next step. Such implementations would provide a rich environment for machine learning, artificial intelligence, and computational photography.

In fact, the environment and concepts of operations will require the evolution of high-end optics and sensors to low- and high-end optics, sensors, and software. These sensors will range from electro-optical, infrared, radar, and any other sensor that might be information for machine learning. There will certainly be large data storage needs. Identification and detection will be done with little information or low-quality information.

While hardware will continue to improve, software capabilities in computational photography, machine learning, and artificial intelligence will advance the capabilities of these systems. Even older or less powerful hardware will excel with the advanced image processing capabilities.

1. REFERENCES

(Note: Use IEEE style)

[1]https://www.hackerearth.com/challenges/hiring/esri-data-science-challenge-2019/.

[2]https://medium.com/data-from-the-trenches/object-detection-with-deep-learning-on-aerial-imagery-2465078db8a9

[3]Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.