Course Work

The use of upscalling and denoising methods to make object recognition in analog type videos better

by:

Gvidonas Pupelis

Coordinator
Vytautas Valaitis

of the

Faculty of Mathematics and Informatics,
Vilnius University

September 4, 2024

Contents

Chapter 1

Introduction

1.1 Context and Importance

In recent years, advancements in deep learning and computer vision have significantly enhanced our ability to process and analyze visual data. One of the key areas where these advancements have been particularly impactful is in video and image processing. The ability to enhance low-quality video footage, for example from drone footage taken using analog type format and then converted to digital format.

1.2 Problem Statement

While modern AI-based models, such as YOLO for object detection and ESRGAN for image upscaling, have made significant strides in digital video processing, they are primarily optimized for high-resolution, well-structured digital footage. These models often struggle when applied to analog-type videos, which present unique challenges, such as:

Noise and Artifacts: Analog videos often suffer from random noise (Gaussian, salt-and-pepper) and artifacts like ghosting and color bleeding, which degrade the quality of the video and make object recognition difficult.

Lower Resolution: Analog formats like NTSC and PAL provide a lower resolution compared to modern digital standards, leading to loss of detail and clarity in object detection.

Color Fidelity Issues: Analog color encoding methods can introduce inaccuracies, making it difficult for algorithms to distinguish between objects and their surroundings.

Despite the advancements in video processing, there is a lack of research and model development specifically aimed at handling the challenges posed by analog video formats. Existing models like YOLO and ESRGAN are trained on high-quality digital datasets,

meaning their performance significantly deteriorates in the presence of analog artifacts, noise, and lower resolutions. This study seeks to explore the limitations of current models when applied to analog-type video, ultimately proving the need for the development of analog-specific AI models that can effectively handle the characteristics of older and noisier video formats.

1.3 Research Objective

The objective of this research is to evaluate the performance and limitations of current AI-based models, specifically YOLO for object detection and ESRGAN for image upscaling, when applied to analog video footage. By working with videos that exhibit the inherent challenges of analog formats—such as noise, low resolution, and common artifacts like ghosting and color bleeding—this study aims to demonstrate the shortcomings of existing models, which are predominantly trained and optimized for high-quality digital data. Through this evaluation, we seek to highlight the need for the development of specialized AI models tailored to the characteristics of analog video. These analog-specific models would address the noise patterns, resolution constraints, and artifacting issues that are unique to analog footage, thus enhancing object recognition and video enhancement in these contexts.

1.4 Significance of the Study

This research addresses a critical gap in AI model performance when applied to analog video, which remains underexplored. Current models like YOLO and ESRGAN are optimized for digital video and struggle with analog-specific challenges such as noise, low resolution, and artifacts. By demonstrating these limitations, this study advocates for the development of specialized AI models tailored to analog formats, improving video processing in fields such as drone racing, surveillance, and archival footage. The findings aim to enhance object recognition and video quality in these applications, ultimately advancing AI's capability to handle diverse video content.

Chapter 2

Literature Review

2.1 Analog Video Characteristics

Digital video formats offer several advantages over analog, including significantly higher resolutions. With the development of HDTV, and later 4K and 8K resolutions, digital video provides much clearer, sharper images than the older analog formats like NTSC and PAL. In addition, digital formats benefit from advanced compression algorithms such as MPEG, H.264, and H.265 (HEVC), which allow for more efficient storage and transmission of video without consuming excessive bandwidth. Furthermore, the post-production process for digital video is much simpler, as it can be edited and manipulated with no loss in quality, unlike analog, where generational loss occurs with each copy or edit . (1)

Analog video formats have been widely used across various applications, but in recent years, they have largely been replaced by digital formats due to the clear advantages digital systems provide. However, analog video still holds an important role in specific contexts, such as drone racing, where the trade-off between video quality and latency becomes critical. (1)

In this work, I will focus on the analog videos used in drone racing. While digital formats have largely taken over due to their superior performance in terms of resolution and flexibility, analog formats are still favored in this high-speed sport because of their lower latency. This lower latency is essential for providing pilots with real-time feedback during races, where even a slight delay could lead to a crash as such it is probable that it would would continue to be a mainstay in this sport.

the drone racing community continues to rely on analog formats due to their minimal latency. In a fast-paced environment where reaction time is crucial, pilots prefer the instantaneous feedback provided by analog video, even if it means sacrificing image quality. Among the most common analog formats used in drone racing are NTSC and PAL.(2)

NTSC (National Television System Committee) is widely used in North America and in FPV (First-Person View) drone racing. Though it offers lower resolution and is limited to standard definition, its low latency makes it ideal for high-speed drone maneuvers . On the other hand, PAL (Phase Alternating Line), more common in Europe and other regions, provides slightly better vertical resolution compared to NTSC. However, this comes at the cost of a marginally higher latency, which can affect performance in high-speed situations like drone racing .

In summary, while analog video formats such as NTSC and PAL lack the image quality, compression, and flexibility of modern digital formats, their critical advantage in latency makes them indispensable in competitive drone racing, where split-second decisions can be the difference between winning and crashing.

2.2 Technical Characteristics and Challenges

Analog video remains a popular choice in drone racing due to its minimal latency, offering near-instantaneous feedback crucial for high-speed maneuvers. However, it comes with challenges like low resolution in formats like NTSC and PAL, which complicates object recognition and visual analysis. Analog signals are also susceptible to interference, noise, and artifacts, such as ghosting and signal dropouts, particularly in complex environments. Additionally, color encoding limitations often lead to color bleeding, further obscuring details. Techniques like upscaling and denoising can help mitigate these issues and hopefully overcome the limitations

2.3 Noise

Gaussian noise, also known as white noise, introduces grainy variations in brightness or color across the video, typically resulting from electronic interference in cameras or transmission systems. In drone racing, poor-quality cameras or receivers can exacerbate this, reducing video clarity and making object recognition difficult, particularly in low light.

Salt-and-pepper noise, manifesting as random black and white pixels, often occurs due to transmission errors or analog-to-digital conversion issues. In drone racing, it obstructs critical details in video feeds, further complicating object detection. (3)

2.4 Artifacts

Ghosting is the appearance of a faint duplicate image, typically caused by multipath interference where signals reflect off surfaces, such as buildings or the ground. This issue, prevalent in environments with obstacles, complicates video analysis in drone racing by introducing confusion into object recognition systems.

Similarly, color bleeding occurs when colors spread into adjacent areas, leading to inaccurate representation. This artifact is often due to limitations in analog color encoding, particularly in NTSC formats, or poor signal quality in FPV systems. It reduces the precision of object edges, making object recognition more difficult.

2.5 Denoising Methods

A range of spatial denoising techniques can be applied to improve the quality of video by reducing noise while maintaining important details. Each method has its strengths depending on the type of noise and the desired level of precision.

Gaussian Blur is a popular method for smoothing images by applying a Gaussian filter, which effectively reduces high-frequency noise like Gaussian noise. However, it can also blur fine details, making it less suitable for applications where sharpness and precision are critical. This method is commonly used in cases where random noise needs to be reduced, and the loss of fine details is acceptable. (4)

Median Filtering, on the other hand, is a non-linear technique that is particularly effective against salt-and-pepper noise. Instead of averaging the surrounding pixels, it replaces each pixel with the median value, preserving edges better than linear filters like Gaussian blur. This makes it ideal for videos suffering from discrete noise spots, such as those found in drone racing footage where maintaining object boundaries is important. (5)

BM3D (Block-Matching and 3D Filtering), while not only spatial but also temporal, is the technique we will adopt in this study due to its advanced capabilities in handling various noise types while preserving important details and edges. BM3D works by grouping similar blocks of pixels within the image, applying collaborative filtering in a 3D transform domain, and reconstructing the image with enhanced clarity. Its ability to maintain sharp edges while effectively reducing noise makes it ideal for high-speed video scenarios like drone racing, where precise object outlines are critical. BM3D stands out as one of the most advanced method in our toolkit for balancing noise reduction and detail preservation. (6)

2.6 upscalling methods

There are numerous AI-based super-resolution techniques available, but two stand out as particularly relevant for improving analog video quality: ESRGAN and SRCNN.

ESRGAN (Enhanced Super-Resolution Generative Adversarial Network) is a highly advanced method that builds upon the earlier SRGAN model. ESRGAN improves on SRGAN by introducing Residual-in-Residual Dense Blocks (RRDB) and a perceptual loss function, which better preserves fine textures and intricate details during the upscaling process. This makes ESRGAN especially well-suited for situations where high-resolution, detail-rich images are critical. In the context of analog video, which often suffers from low resolution and noise, ESRGAN's ability to enhance fine details and produce high-quality results is invaluable. This technique is particularly useful in restoring and enhancing degraded analog footage, making it a perfect candidate for modernizing analog video, such as in drone racing, where clarity and detail are essential for object recognition. While ESRGAN is computationally intensive, advances in hardware and optimization have made real-time applications, such as live video feeds in drone racing, increasingly feasible.(7)

SRCNN (Super-Resolution Convolutional Neural Network), on the other hand, is one of the earliest AI-based approaches to super-resolution. Using deep convolutional neural networks, SRCNN learns the mapping from low-resolution to high-resolution images. While simpler and less computationally demanding than ESRGAN, SRCNN is still an efficient method for enhancing analog video. Its straightforward architecture allows for real-time application in competitive drone racing, where low latency is crucial. Though SRCNN might not match ESRGAN's ability to preserve very fine details, it remains a versatile and effective approach, particularly when computational resources are limited. (8)

In this study, we will use ESRGAN as it offers the best balance between detail preservation and image enhancement, making it the ideal choice for improving the quality of analog video footage in dynamic environments such as drone racing.

2.7 Object Recognition Models

Object recognition models are widely used for detecting and classifying objects in various environments, including low-quality or noisy video footage. Two of the most prominent approaches include You Only Look Once (YOLO) and Region-Based Convolutional Neural Networks (R-CNNs), each offering different strengths depending on the nature of the task.

YOLO models are designed to balance speed and accuracy, making them highly efficient for object detection by dividing the image into a grid and predicting bounding boxes and probabilities for each grid cell. YOLO's design allows for rapid detection, which

is particularly useful in scenarios where real-time processing is not strictly required but efficiency is still a key factor. However, when applied to low-quality, noisy videos, such as those encountered in analog formats, YOLO's grid-based approach can struggle to capture smaller details and objects. Noise can interfere with pixel-level predictions, leading to inaccuracies. Despite these challenges, YOLO remains a robust option due to its efficiency and versatility in handling various object recognition tasks, even in less-than-ideal video conditions. (9)

In contrast, Region-Based Convolutional Neural Networks (R-CNNs)—which include R-CNN, Fast R-CNN, and Faster R-CNN—operate by first proposing regions of interest (ROIs) and then applying a CNN to classify objects within these regions. While R-CNN models are known for their high accuracy and ability to detect objects in detail, they face challenges in low-resolution and noisy video environments. In such cases, the proposed ROIs may lack the clarity needed to accurately classify objects, leading to increased false positive and false negative rates. Additionally, noise can interfere with the selection of ROIs, further reducing accuracy. Although R-CNN models offer superior accuracy in high-quality video, their sensitivity to noise and lower performance in degraded video makes them less practical for use in environments where image quality is compromised. (10)

For this study, YOLO will be used due to its balance between efficiency and robustness, making it an appropriate model for object recognition in noisy and low-resolution video footage. While real-time performance is not the primary focus, YOLO's overall efficiency allows it to handle object detection tasks even in challenging environments where video quality is suboptimal.

2.8 Gap in literature

While significant progress has been made in the development of AI models for video processing these models are primarily trained and optimized for high-quality digital footage. The literature predominantly focuses on enhancing performance in environments with well-structured, high-resolution data, leaving a gap in addressing the specific challenges posed by analog video formats. Analog videos, characterized by issues like noise, low resolution, and artifacts such as ghosting and color bleeding, remain underexplored in the context of AI-based enhancement and object recognition. Current studies focus on improving performance in digital contexts, with limited attention given to how these models perform on analog data, despite its continued use in fields such as drone racing, surveillance, and archival footage. This gap highlights the need for research that not only investigates how existing AI models struggle with analog video but also emphasizes the development of analog-specific models that can effectively address these challenges.

Chapter 3

Methodology

3.1 Data Collection

The video data used in this study was captured during a drone racing event, where the focus was not initially on object recognition or footage for this specific research. The footage is analog, recorded in an outdoor environment on a grassy field, featuring various natural and artificial obstacles. While the primary focus of the drone race was navigating around obstacles, the video also captures individuals present in the area. These individuals, who are primarily spectators or race staff, serve as the primary objects of interest for the object recognition task in this study.

The footage is recorded in MP4 format at a resolution of 720x576 pixels and runs at 50 frames per second. For the purposes of processing and object recognition, we extract a frame every 25 frames, equating to one frame every 0.5 seconds. The objective is to recognize and detect individuals in these frames despite the challenges presented by analog video formats, such as noise, glare from the sun, and occasional low resolution.

3.2 Denoising Process

For denoising, we exclusively use BM3D (Block-Matching and 3D Filtering), as initial experiments revealed that Gaussian blur caused excessive loss of image detail, making the footage difficult to interpret even for human observers. BM3D, on the other hand, is a highly effective denoising method that preserves fine details while reducing noise, making it suitable for this study. This process helps mitigate the noise and artifacts commonly found in analog footage, ensuring that important visual information remains intact for object recognition.

3.3 Upscaling Process

To enhance the resolution of the extracted frames, we apply Real-ESRGAN, a pre-trained model based on the Enhanced Super-Resolution Generative Adversarial Network. While the model is not specifically tailored for analog video, it is capable of significantly improving video clarity and detail. The upscaling process is intended to improve the quality of the video, making it easier to detect and recognize individuals and objects despite the analog video's limitations.

3.4 object recognition

For object recognition in the video frames, we employed a pre-trained YOLO (You Only Look Once) model, known for its speed and accuracy in detecting objects within images. The video footage was processed by extracting one frame every 25 frames, and each frame was resized and normalized to fit YOLO's input requirements. The model predicts bounding boxes and class labels for objects like people and obstacles present in the footage. However, due to the limitations of the analog video, including noise and low resolution, YOLO struggled to detect objects accurately, yielding almost no confirmed positives in the results.

Chapter 4

Results

4.1 Denoising

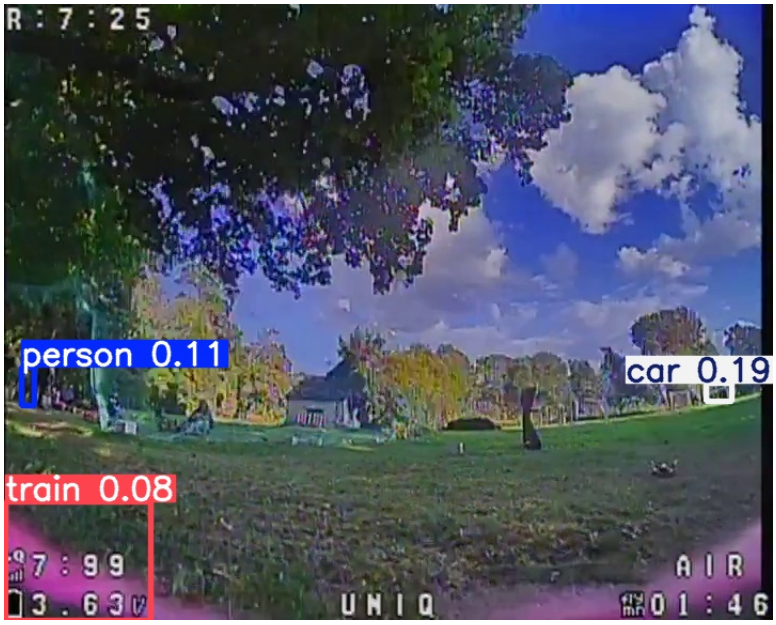
The first step in processing the video frames was to reduce the noise using the BM3D algorithm, applied to each of the RGB channels separately. We used a sigma value of 0.05 and applied the algorithm three times for each channel to amplify the denoising effect. While the visual improvements were subtle and not immediately obvious to the naked eye, they were present and measurable upon closer inspection.

To objectively assess the impact of the denoising, we calculated the Peak Signal-to-Noise Ratio (PSNR) for the images before and after the BM3D process. On average, the denoised images yielded a PSNR value of approximately 30 dB, which indicates a moderate change in the image. Images that had lower noise showed slightly higher PSNR values, around 32 dB, suggesting that the BM3D algorithm made fewer adjustments in these cases. Conversely, images with more significant noise saw PSNR values drop to 28 dB, reflecting a greater impact of the denoising process as the algorithm worked harder to remove noise.

4.2 object recognition

We have 4 images here:

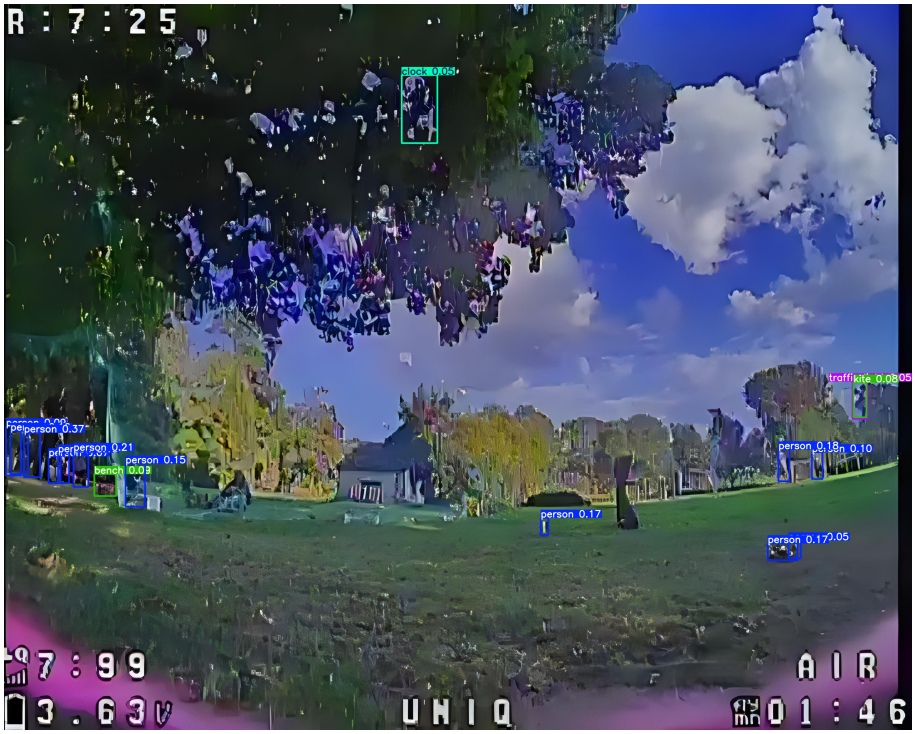
this one has nothing done to it



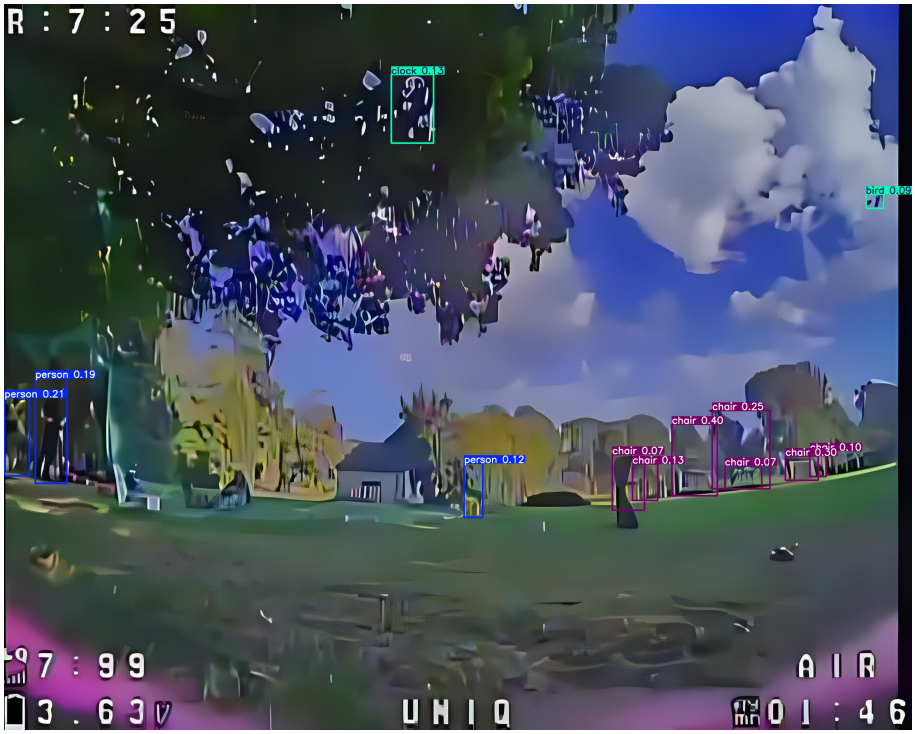
This one has been denoised

As we can see there's almost no improvement and the object recognition sees nothing although our human eyes could pickup the people on the left

this one has been upscalled with ESGRAN without denoising:

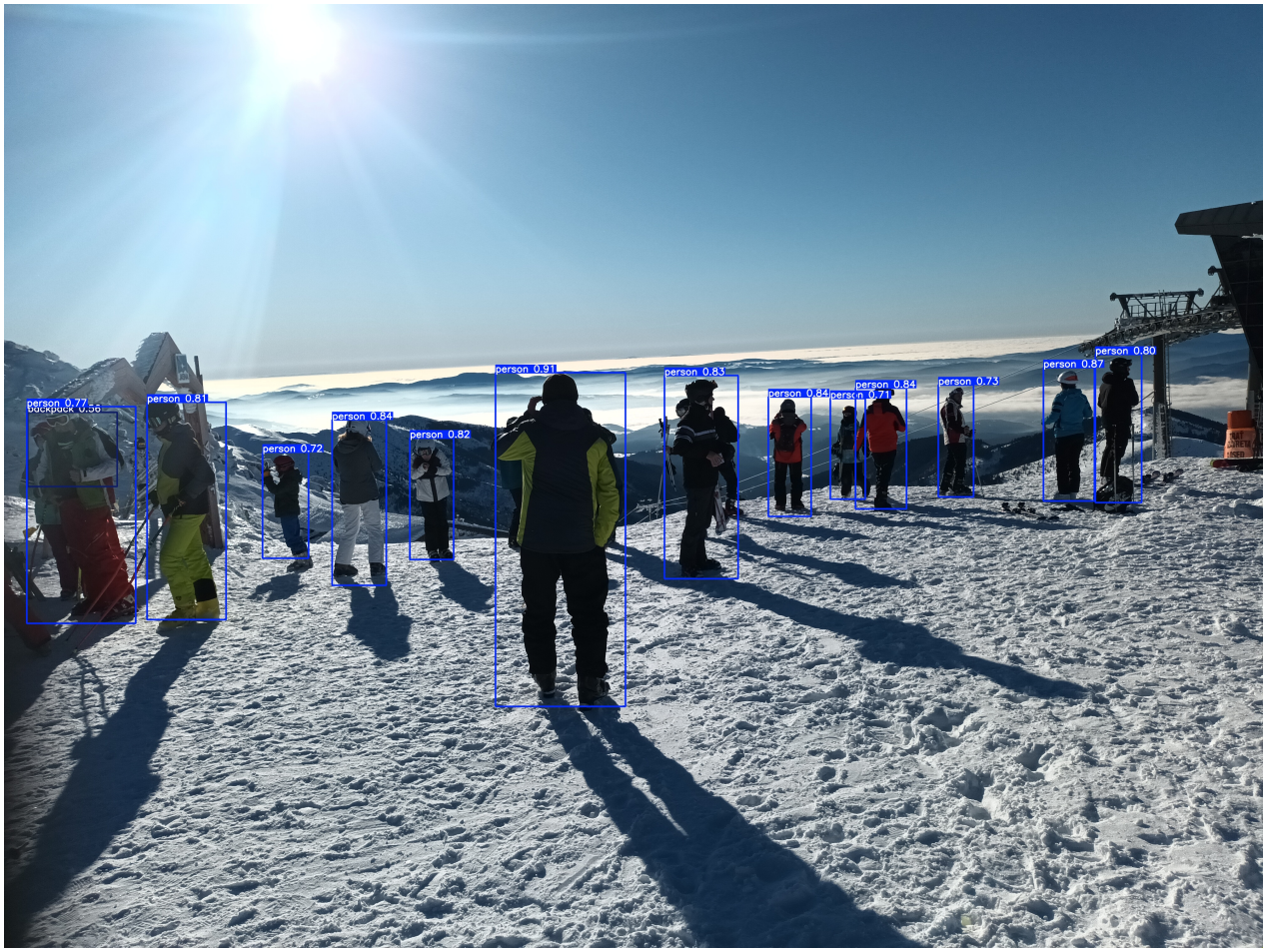


and this one has been denoised with BM3D has been upscalled with ESGRAN



At this point the image is smooth and as such YOLO even manages to tag a person on the left with 21 percent however it also tags a chair that's a house in the distance with 41 percent

And for comparison A normal digital image that was taken recently:



It spots every single person without fail

Chapter 5

Discussion

The results of our experiments highlight the challenges of applying modern AI models, such as YOLO and Real-ESRGAN, to analog video footage. Despite using advanced denoising techniques like BM3D and upscaling with Real-ESRGAN, the analog video frames failed to produce reliable object recognition results using YOLO. This was expected, as analog videos are inherently prone to noise, low resolution, and visual artifacts like glare, which degrade the quality of the footage and limit the effectiveness of object recognition models that rely on clear, high-resolution data.

Our attempts to denoise the analog footage using BM3D improved the video quality slightly, as evidenced by the modest increase in PSNR values. However, this enhancement was not sufficient to enable YOLO to detect objects, such as people and obstacles, in the video. Even after upscaling the frames with Real-ESRGAN, YOLO failed to generate any confirmed positives, underscoring the limitations of applying digital-centric models to analog video.

Interestingly, when we tested YOLO on a comparable digital image extracted from a video with higher inherent quality, the model performed flawlessly. YOLO successfully recognized and labeled all relevant objects, in the digital footage, with high accuracy and precision. This stark contrast between the performance on analog versus digital video emphasizes the critical role that video quality plays in object recognition tasks. Digital footage, with its higher resolution, absence of noise, and better color fidelity, aligns more closely with the conditions YOLO was trained on, explaining its superior performance.

These results indicate a clear gap in the capability of current AI models to handle analog video formats effectively. While modern AI models like YOLO are optimized for high-resolution, noise-free digital data, they struggle significantly with the challenges inherent in analog formats. The failure to detect objects in analog video—even after significant efforts to denoise and upscale—suggests that there is a need for models that are specifically

trained and optimized for analog video characteristics. Analog-specific models would need to handle the lower resolutions, noise patterns, and artifacts that are unique to these formats, which are still relevant in certain fields such as drone racing, surveillance, and archival footage.

In conclusion, while AI-driven object recognition models like YOLO perform well on high-quality digital data, they are not equipped to handle the complexities of analog footage. This highlights the necessity for future research focused on developing specialized AI models capable of processing and enhancing analog video effectively, ensuring more accurate object recognition and video analysis in these contexts.

Chapter 6

Conclusion

This study has highlighted the significant limitations of existing AI models, such as YOLO and Real-ESRGAN, when applied to analog video footage. Although these models excel in high-quality, digital video environments, their performance degrades drastically in the presence of analog-specific challenges such as noise, low resolution, and artifacts. Despite attempts to mitigate these issues through denoising with BM3D and upscaling with Real-ESRGAN, YOLO failed to accurately detect any objects in the analog footage, whereas its performance on digital video was flawless.

The results strongly indicate a pressing need for the development of AI models specifically tailored to handle the unique characteristics of analog video. Current models are designed primarily for digital data, and their inability to manage the complexities of analog formats demonstrates the gap in AI capabilities for this type of media. Moving forward, the creation of analog-specific models, which are trained to account for the lower resolution, noise, and artifacts inherent in analog video, is necessary. Such models would better handle analog footage, offering improved object recognition and video enhancement in fields where analog video remains in use, such as drone racing, surveillance, and archival footage.

In conclusion, this study underlines the need for new AI model creation focused on processing and enhancing analog video formats. Future research should prioritize this gap, ensuring that AI can be effectively applied to analog media, thus bridging the divide between digital and analog video processing capabilities.

Chapter 7

References

1. Drone Nodes. (2020). FPV Analog vs. Digital Systems: What's Best for Drone Racing? Retrieved from Drone Nodes. Here

2. All About Analog FPV Video and the ClearView's Magic. Retrieved from GetFPV. Here

3.Image Denoising Review: From Classical to State-of-the-Art Approaches" — Information Fusion, Elsevier Here

4.Gonzalez, R.C., Woods, R.E., Digital Image Processing. Here

5. Leiou Wang "A New Fast Median Filtering Algorithm". Here

6. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K. (2007).Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. Here

7. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Change Loy, C. (2018). "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks". Here

8. Dong, C., Loy, C.C., He, K., Tang, X. (2015). "Image Super-Resolution Using Deep Convolutional Networks". Here

9. Redmon, J., Farhadi, A. (2016). "YOLO9000: Better, Faster, Stronger". Here

10. Ren, S., He, K., Girshick, R., Sun, J. (2015). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks" Here