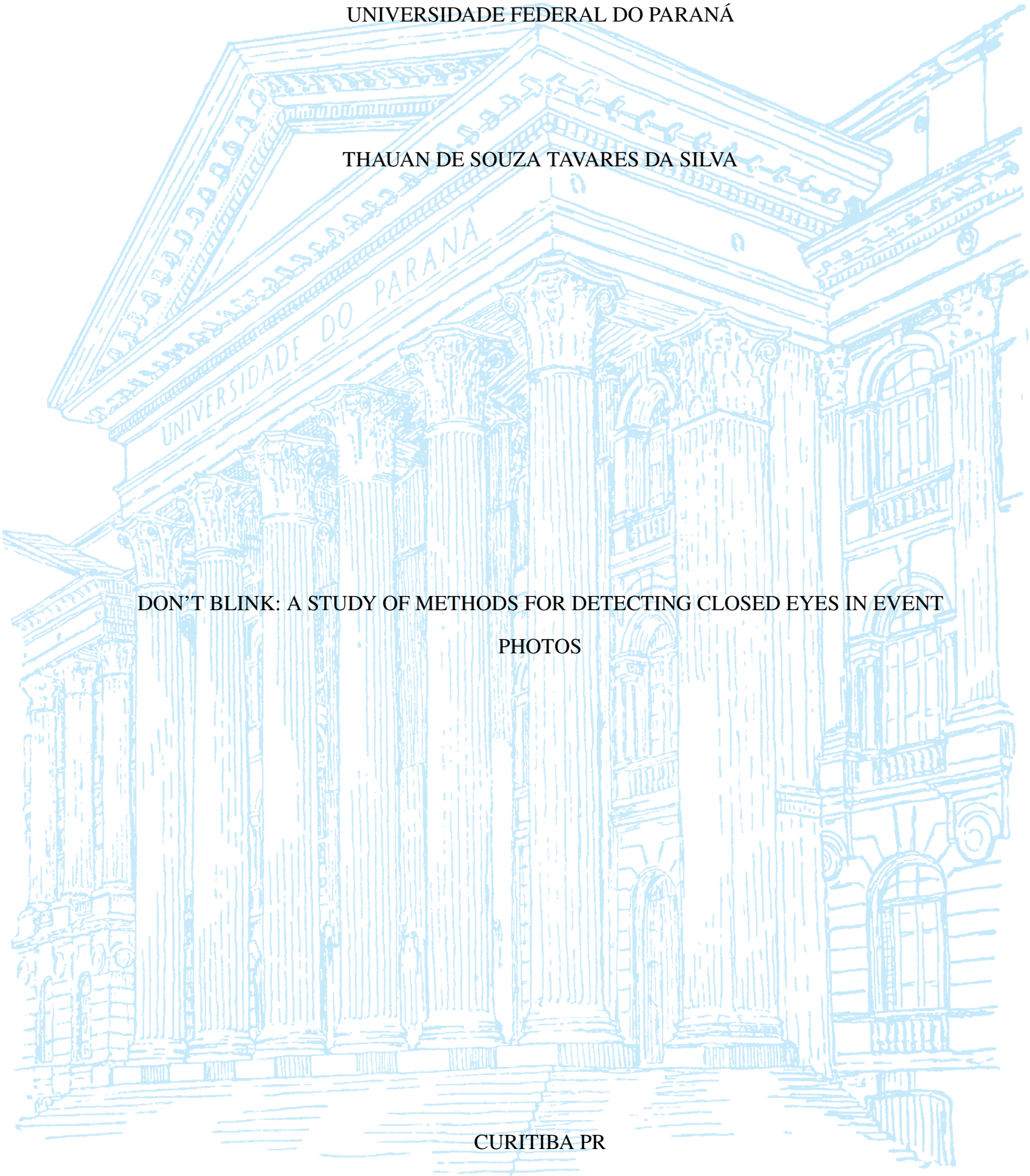
UNIVERSIDADE FEDERAL DO PARANÁ

THAUAN DE SOUZA TAVARES DA SILVA



DON’T BLINK: A STUDY OF METHODS FOR DETECTING CLOSED EYES IN EVENT PHOTOS

CURITIBA PR

2024

THAUAN DE SOUZA TAVARES DA SILVA

DON'T BLINK: A STUDY OF METHODS FOR DETECTING CLOSED EYES IN EVENT PHOTOS

Trabalho apresentado como requisito parcial para a conclusão do curso de Bacharelado em Ciência da Computação, no Departamento de Ciências Exatas, Universidade Federal do Paraná..

Área de concentração: *Computação*.

Orientador: David Menotti.

CURITIBA PR

2024

Universidade Federal do Paraná
Setor de Ciências Exatas
Curso de Ciência da Computação

Ata de Apresentação de Trabalho de Graduação II

Título do Trabalho: ***DON'T BLINK: A STUDY OF METHODS FOR DETECTING CLOSED EYES IN EVENT PHOTOS***

Autor(es):
GRR 20171591 None: THAUAN DE SOUZA TAVARES DA SILVA
GRR _____ Nome: _____
GRR _____ Nome: _____

Apresentação: Data: 13 / 12 / 2024 Hora: 08:00 Local: https://meet.google.com/efc-ajdu-fwg

Orientador: DAVID MENOTTI GOMES
Membro 1: BRUNO H. KAMAROWSKI DE CARVALHO
Membro 2: GABRIEL EDUARDO LIMA
(nome) (assinatura)

AVALIAÇÃO – Produto escrito		ORIENTADOR	MEMBRO 1	MEMBRO 2	MÉDIA
Conteúdo	(00-40)				30
Referência Bibliográfica	(00-10)				09
Formato	(00-05)				03
AVALIAÇÃO – Apresentação Oral					
Domínio do Assunto	(00-15)				15
Desenvolvimento do Assunto	(00-05)				05
Técnica de Apresentação	(00-03)				02
Uso do Tempo	(00-02)				01
AVALIAÇÃO – Desenvolvimento					
Nota do Orientador	(00-20)		**********	**********	20
NOTA FINAL		**********	**********	**********	85

Pesos indicados são sugestões.
Conforme decisão do colegiado do curso de Ciência da Computação, a entrega dos documentos comprobatório de trabalho de graduação 2 deve deve respeitar os seguintes procedimentos: Orientador deve abrir um processo no Sistema Eletrônico de Informações (SEI – UFPR); Selecionar o tipo: Graduação: Trabalho Conclusão de Curso; informar os interessados: nome do aluno e o nome do orientador; anexar esta ata escaneada e a versão final do pdf da monografia do aluno.; Tramita o processo para CCOMP (Coordenação Ciência da Computação).

To Allana and Chloe...

ACKNOWLEDGEMENTS

First and foremost, I thank God for being my source of strength, wisdom, and guidance in all moments of my life. To my wife, Allana, who has always been by my side with love, patience, and unconditional support, giving me strength in difficult times and sharing in my joys.

To my parents, Josué and Silvia, who taught me the most important values in life, always encouraging me to chase my dreams and become a better person every day. To my sister Yanajara, Fabinho, Yuri, Thales, and Kiara, for always being there, supporting me, and motivating me to keep moving forward.

To father and mother-in-law, Josuel and Wilma, and my little brother-in-law Pedro, for their warmth, care, and constant support, always with love and generosity. To my entire family, who has always inspired me and provided the support needed to continue with perseverance and faith in my path.

To my professor, David Menotti, for his guidance, dedication, and patience, not only in the development of this thesis but also throughout the entire course. His guidance and help were essential to my academic and personal growth.

Lastly, but by no means least, I am immensely grateful to my beloved daughter, Chloe, who has arrived to complete my happiness and teach me every day the true meaning of love.

To all of you, my sincere gratitude.

ABSTRACT

Photography has evolved into a professional and artistic field, and with the rise of digital technology and social media, the volume of photographs has significantly increased. This has created new challenges for photographers, especially in the curation process. This study investigates the automation of closed-eye detection in event photographs, an important but time-consuming task in photo selection. The research focuses on comparing different versions of the YOLO (You Only Look Once) model for their effectiveness in detecting closed eyes, with an emphasis on YOLOv8 and YOLOv11. By evaluating these models, this work aims to provide a detailed comparison of their performance in terms of precision and practical applicability in professional photography workflows. While a fully automated photo curation system is beyond the scope of this work, the study serves as a foundational step towards developing such tools. The findings contribute to the field by providing insights into how object detection models, particularly YOLO, can be applied to streamline the photo selection process, reducing manual effort and human error while ensuring high-quality photo curation. Additionally, the research offers a proprietary database of annotated event images that can support future advancements in automated photographic curation.

Keywords: eye detection. image process. photography.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

DINF	Departamento de Informática
PPGINF	Programa de Pós-Graduação em Informática
UFPR	Universidade Federal do Paraná
OBB	Oriented Bounding Box
YOLO	You Only Look Once
mAP	Mean Average Precision
mAP50	Mean Average Precision at IoU (Intersection over Union) threshold 50%
mAP50-95	Mean Average Precision at IoU thresholds from 50% to 95%
P	Precision
R	Recall
TP	True Positives
FP	False Positives
IoU	Intersection over Union

CONTENTS

1 INTRODUCTION

As a means of documentation or a commercial tool, photography plays a central role in contemporary society. The popularization of digital cameras and smartphones, combined with social media, has democratized the act of capturing images, generating a massive volume of photographs daily. According to a study published by Repsly as illustrated in Figure 1.1. it is estimated that globally, millions of photos are captured every day, as shown in the image.

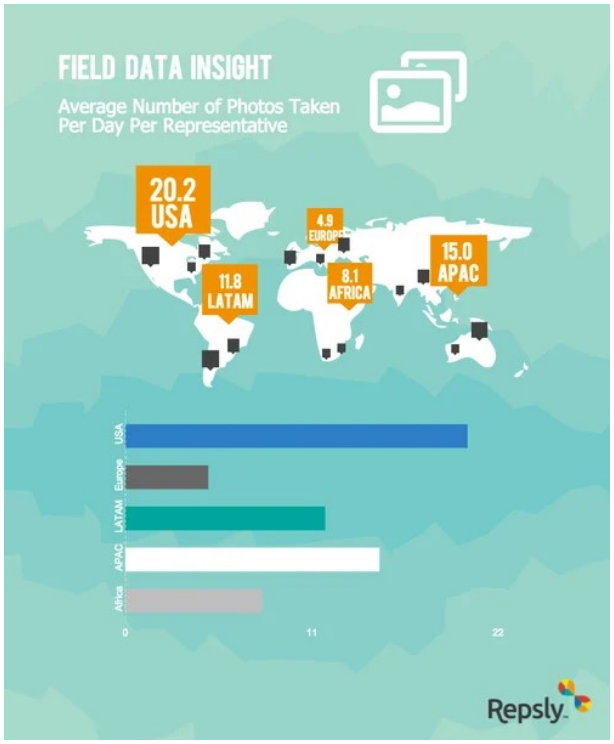


Figure 1.1: Illustration of the Research Conducted by Repsly. (Repsly, 2024)

In the professional context, photographers must balance technical expertise and creativity to meet their clients' expectations for quality and visual storytelling. This scenario presents significant challenges, especially in the curation process, which requires time and effort to select the best photos from thousands of captures. In events such as weddings and birthdays, the need to capture spontaneous and emotive moments often results in photographs being discarded due to technical issues such as closed eyes, blur, or poor exposure. Therefore, tools capable of automating part of this process are increasingly necessary to enhance efficiency and reduce human errors.

This study focuses on applying computer vision techniques, particularly object detection models from the YOLO (You Only Look Once) family, to automatically identify photos where people have their eyes closed. This approach aims not only to optimize image selection but also to provide technical support to photographers, allowing them to concentrate their efforts on more complex and subjective criteria, such as artistic composition and visual narrative.

In addition to addressing a practical need in the photography market, this study contributes to the field of machine learning applied to photography by offering a detailed comparative analysis of different YOLO versions and creating an annotated dataset that can be used for future research. The main goal is to explore the boundaries and possibilities of these

technologies to automate stages of photographers' workflows, saving time and ensuring a high standard of quality.

1.1 MOTIVATION

Photo curation is one of the most critical and labor-intensive steps in the workflow of professional photographers. In a typical event, taking a wedding as an example, thousands of photos may be captured, of which only a fraction will be selected for editing and delivery to the client. This process not only consumes time but is also prone to errors caused by fatigue or time constraints.

One recurring issue in this process is identifying photos where people have their eyes closed, something that often happens during spontaneous moments or due to involuntary blinking. Non-intentional photos with closed eyes, although technically correct in other aspects, are generally discarded because they do not meet the client's aesthetic expectations. This can be illustrated by Table 1.1, which shows that, in a typical wedding in Brazil with two photographers, approximately 3,000 to 4,000 photos are taken, of which only 20% to 30% are selected for editing, while the rest are discarded.

	Making of	Ceremony	Photoshoot	Wedding Party
Photos taken	600	800	300	1500
Discarded Photos	400	700	250	800
Utilization	33,3%	12,5%	16,6%	46,6%

Table 1.1: Average Use of photos at a wedding

Additionally, the workflow of a wedding in Brazil usually follows a standardized structure composed of several stages, as illustrated in Figure 1.2. This workflow includes the bride and groom's "getting ready" session, the religious or civil ceremony, the couple's photoshoot after the ceremony, their entrance into the reception hall, photos with guests, and special moments during the party. Photographs are continuously taken in all these phases, resulting in a large volume of images, many of which are discarded due to technical issues such as closed eyes.

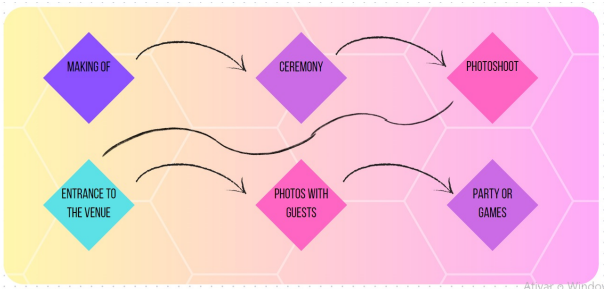


Figure 1.2: Example of a Wedding Photography Workflow. Source (The author).

Automated closed-eye detection can significantly help reduce this volume of discarded photos and expedite the curation process. The motivation for this study lies in the potential to create tools that can transform this specific aspect of photo curation. By automating the identification of problematic photos, photographers can dedicate more time to creativity and client engagement while reducing the manual effort required to review large volumes of images.

Furthermore, this work aims to pave the way for the application of artificial intelligence in creative tasks, demonstrating how computer vision models can be adapted and applied to fields traditionally dependent on human decision-making. By creating a proprietary dataset

and utilizing advanced versions of the YOLO model, this study not only explores the technical feasibility of closed-eye detection but also contributes to the evolution of curation technologies in the photography market. The expected impact is significant time and effort savings, enabling photographers to deliver high-quality results more quickly and efficiently.

1.2 OBJECTIVES

This work aims to study and implement techniques for detecting closed eyes in event photographs, to automate part of the photo-curation process. As a result of this work, a detailed study and supporting resources are provided to aid in the development of solutions capable of automatically identifying photos where all subjects in the scene have their eyes open. While a completely automated photo curation tool requires a more comprehensive study, this work specifically focuses on the detection of closed eyes as a first step toward this automation. This study serves as a foundation for future implementations of tools that assist photographers in selecting high-quality photos.

1.3 CONTRIBUTION

This work contributes to the field of photography by proposing a study to photo curation through the automation of processes focused on detecting closed and open eyes. The contributions are
Using techniques for detecting open eyes:

- Study of techniques and algorithms for detecting open and closed eyes and implementation of techniques that allow for image analysis and separate images where the eyes are closed and open.

Image Database:

- A proprietary photo database, which will be made public, containing complete photoshoots and events available in high-quality "raw" format.
- The dataset (Tavares, 2024) can be accessed through the following link: Google Drive - DontBlinkDataset.

In addition to enhancing the efficiency of photographers' work, this study also advances research that uses artificial intelligence in creative and artistic tasks. The techniques developed and the proposed approach have the potential to transform the way photo curation is conducted, offering a practical solution to a common problem in the field of photography.

1.4 DOCUMENT STRUCTURE

The remainder of this document is organized as follows. Chapter 2 presents the background of this work, discussing the photography market and exploring technical aspects from both the computational and photographic perspectives. Chapter 3 reviews the related works, providing detailed descriptions of some of the analyzed studies that are related to this work. Chapter 4 describes the materials and methods used in this study, including the creation of a proprietary database and the configurations of the YOLO models analyzed. Chapter 5 presents the results obtained, along with their interpretation and analysis, highlighting the performance of the proposed approaches.Finally, Chapter 6 concludes the work, summarizing the main findings, discussing limitations, and suggesting directions for future research

2 BACKGROUND

This chapter provides an essential foundation for understanding the context and relevance of this work. Here, we explore the photography market, highlighting current trends and the importance of photo curation in the professional landscape. Additionally, we discuss fundamental technical aspects from both computational and photographic perspectives, which are crucial for developing effective methodologies for image analysis and selection.

2.1 MARKET

As we have seen, photography has become a significant market in our time, with much of this popularity attributed to the advancement and popularization of social media. Using Brazil as a case study, a country with a large population of over 200 million inhabitants and a traditional love for parties, it is common today to have gender reveal parties, monthly celebrations, birthday parties, weddings, graduations, and a variety of other events, often employing professional photographers. Focusing on just one segment of the market, the wedding industry, according to a study conducted by the Brazilian Association of Events (CBN, 2023) in 2023, the wedding market in Brazil has generated 40 billion reais per year. Another data point that supports these findings is from the Civil Registry Transparency Portal, which shows in Brazil 914,500 weddings were registered in 2023.

Another major financial driver heavily influenced by photography is the digital camera market, which is crucial for photographers as these are their tools of the trade, as shown by Mordor Intelligence (Intelligence, 2024). Despite a general decline in the digital camera market, it still generates about $5.39 billion. This decline is because the average consumer prefers to invest in a smartphone with a good camera, as it is more affordable, especially in Brazil. In response, major brands like Canon and Nikon have developed full-frame cameras for a niche of clients who seek faster shutters, better resolution, and greater sharpness. These clients are mostly professionals, such as sports and wedding photographers, whose livelihoods depend on photography, and a poor-quality photo could negatively impact their business.

There are estimates that the camera market is growing after a decline due to the preference for smartphones and a significant drop during the Covid-19 pandemic in 2020, which directly affected the events market. With the prohibition of events and travel, the camera market was significantly impacted. However, post-pandemic, the market is rapidly growing, and camera companies have shown growth due to the release of new products, such as Mirrorless cameras. Nikon, for example, reported revenue of 64.3 billion yen (US$ 460.3 million) in the first two quarters of the 2022 fiscal year, due to an expansion in the sales mix with a shift to professional/hobby models aided by Mirrorless camera sales, as illustrated in Figure 2.1.

Given this context, we can see how a tool that helps optimize editing time would be important to further boost this market. Such a tool would allow photographers to dedicate more time to other aspects of their work, such as client service, marketing to attract new clients, and also contribute to increasing their leisure time.

2.2 PHOTOGRAPHY

Taking professional photos, as discussed, requires extensive study by the photographer, as techniques have been developed over the years to capture images properly and express the

Figure 2.1: Mordor Intelligence Research & Advisory. (2023, June). Tamanho do mercado de câmeras digitais e análise de participação – Tendências e previsões de crescimento (2024 – 2029). Mordor Intelligence. Retrieved June 18, 2024, from https://www.mordorintelligence.com/pt/industry-reports/digital-camera-market

photographer's vision in each image. There are numerous techniques that help take good photos. Light techniques like the Bokeh effect, which comes from the Japanese "Boke" meaning blur in English, is a technique that blurs the background with the entry of light. Motion capture techniques such as Panning, require a slow shutter speed and concentration to follow the subject of interest at the same speed, making the subject appear still while the background conveys a sense of speed. Framing techniques like the Rule of Thirds, where the photographer mentally divides the image into 9 equal parts and centers the subject at the intersection points. These techniques are crucial for ensuring good image quality and help in selecting images that can be considered good.

In addition to these techniques used to capture images, there are others that help in curating these images for editing. Some images are easy to assess for quality by looking at the exposure; typically, images with high or low exposure are not used because they are difficult to recover unless the photographer intends to convey a specific emotion or perception. However, such images are often taken due to changes in lighting at events, resulting in adjustment shots that are discarded. Another important detail for photo selection, which is easier to detect, is people with their eyes closed in the scene. Unless the intention is to capture people with closed eyes, these photos occur when the subject blinks at an unexpected moment. Therefore, professionals often take multiple shots of the same scene.

Another factor that is a bit more difficult to perceive for people who are not in the habit of selecting photos is whether the objects of interest are properly focused. In photographers' jargon, we say the photo is "sharp." Determining this is difficult because some photos may appear focused, but when opened on a computer for editing and zoomed in, it becomes clear that they are not well-focused. Some photos are easier to identify as out of focus if the camera's shutter did not open properly due to a mechanical failure, resulting in a completely blurry, undefined image. Therefore, training a neural network or using algorithms to automate photo selection is feasible, as we can replicate and teach these techniques to a computer.

2.3 ALGORITHMS AND TECHNOLOGIES

2.3.1 Image Processing

The history of image processing is marked by significant advances and important discoveries that have shaped the field over the decades. In the 1950s and 1960s, image processing began to gain attention with the development of basic algorithms for image analysis. One of the first works to highlight these concepts was the paper by L. W. Seitz and J. F. Riley, published in 1955, titled "Image Processing."(G. Kovasznay and Joseph, 1955). This study discusses the visualization of scalar functions of two independent variables as images and how mathematical operations can be interpreted as modifications or processing of these images, establishing a foundation for future research in the field.

In the 1970s, the development of digital image processing techniques began to take shape, with one of the most notable events involving Alexander Sawchuk. Sawchuk's work included a notable study on image compression using the Lenna image, a famous photograph of a model published in Playboy in 1972. This image was widely used as a standard test bed for evaluating and comparing image processing algorithms, such as filtering and compression. The use of the Lenna image became a landmark in image processing research due to its broad acceptance and reference in various studies and benchmarks in the field.

The 1980s were characterized by advancements in image segmentation. The paper "A Study of Edge Detection Algorithms" (T. Peli, 1982) represents this phase well. This work provided a detailed analysis of edge detection techniques, highlighting methods for identifying and analyzing edges in digital images, which is essential for segmentation and interpretation of images.

In the 1990s, the field of image processing saw significant advancements with the development of real-time processing techniques and the beginning of computer vision. An important example is the paper "Object Recognition from Local Scale-Invariant Features," published in 1999 by David Lowe (Lowe, 1999). This study introduced the SIFT (Scale-Invariant Feature Transform) algorithm, which enabled robust and efficient object recognition in images, revolutionizing how machines interpret and identify visual features. Starting from the 2000s, image processing techniques began to utilize deep learning and neural networks, as discussed below.

2.3.2 Neural Networks

Artificial neural networks have played a fundamental role in the evolution of image processing techniques, especially in contemporary photography. Among the various neural network architectures, Convolutional Neural Networks (CNNs) have stood out for their efficiency in image analysis tasks. Specifically designed to process data in the form of multidimensional matrices, such as images, CNNs are capable of automatically identifying and extracting essential features like edges, textures, shapes, and complex patterns. According to Voinov, in "Deep Learning-based Vessel Detection from Very High and Medium Resolution Optical Satellite Images as Component of Maritime Surveillance Systems."(Voinov, 2020), a CNN is typically structured through the repetition of convolutional, pooling, and fully connected layers, allowing the network to learn and extract hierarchical complex features from images. In the first stage, filters (or kernels) are applied to the input image to detect local features, such as edges, textures, and patterns. Each filter is trained to recognize a specific pattern. The filters traverse the image, performing convolution operations that produce a feature map, where each pixel represents the presence of the pattern detected by the filter. The filter is trained to detect specific features,

ranging from simple structures like edges and curves to more complex ones like ears, eyes and noses as explained in “Machines that can see: Convolutional Neural Networks". (Tejani, 2016)

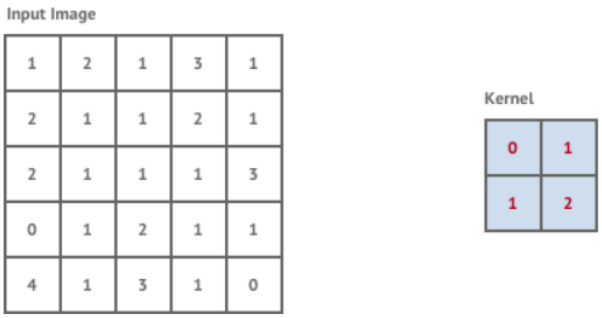


Figure 2.2: Input Image and the Kernel. Source (Tejani, 2016)

After applying the filters, the results are processed through a non-linear activation function, such as ReLU (Rectified Linear Unit), which allows for modeling non-linear relationships between the image features. The result of each filter is the resulting feature maps, which are used as input for the next layer (Dettmers, 2015). Following the typical structure described by Voinov (Voinov, 2020), the next block is the Pooling Layer, where the feature maps are processed by pooling layers. Their function is to reduce the dimensionality of the feature maps generated by the convolutional layers, decreasing the size of the data and computational complexity while preserving the most important features.

Techniques such as Max Pooling (which selects the maximum value in a region) or Average Pooling (which calculates the average of the values) are used to condense the information, reducing computational complexity and preserving features.

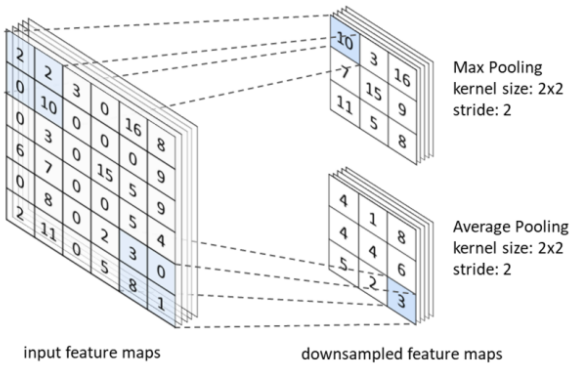


Figure 2.3: feature maps are limited by Max Pooling or Average Pooling. Source (Voinov, 2020)

Finally, we have the third block, which is the Fully-Connected layer, where classification is performed by fully connected layers. In this block, the neurons of one layer are connected to all the neurons of the next layer, transforming the feature maps into a probability vector. This vector represents the probability that the image belongs to certain object classes. The final layer, usually using an activation function, generates the final classification, determining the class to which the image most likely belongs, as exemplified by (Tejani, 2016), in the example of determining whether the image is of a lion or a tiger.

CNNs are not limited to simple image classification; they are applied in a wide range of tasks, including semantic segmentation, object detection, and even action recognition in image sequences. The ability of these networks to decompose images into fundamental components

and process them hierarchically has led to significant advances in areas such as surveillance, vehicle automation, and image-based medical diagnostics.

In his paper on deep learning (LeCun et al., 2015), Lecun stated that in the future, progress in computer vision is expected to come from systems that combine CNNs with recurrent neural networks (RNNs) and reinforcement learning, allowing these networks not only to interpret images passively but also to make intelligent decisions about where to focus their attention in a scene. Since then, CNNs have transformed the way images are processed and interpreted, enabling the automation of complex processes in the curation and management of photographic collections. The efficiency and accuracy provided by the deep learning techniques employed in CNNs have opened up new creative and operational possibilities, driving innovation in various technological applications.

2.3.3 Object Detection - YOLO

Object detection is a technique used to identify and locate different objects in an image, classifying them according to their categories. The process involves balancing the model's speed with accuracy in classification and localization, which is challenging due to variations in scale, rotation, and position of objects in images. There are two main types of detectors: two-stage and one-stage detectors.

In two-stage detectors, Regions of Interest (ROIs) are identified beforehand, increasing accuracy but potentially compromising speed. On the other hand, one-stage detectors process the image more quickly, without the need to pre-select ROIs, but generally with less precision.

Using a one-stage detector, in 2016, a system called "You Only Look Once" (YOLO) was proposed by (Redmon et al., 2016), which transformed the object detection problem into a regression problem, allowing the network to detect to which class an object belongs. The YOLO approach was revolutionary compared to two-stage object detectors due to its ability to perform real-time detection. The YOLO model divides the input image into an SxS grid, and each cell in this grid is responsible for predicting bounding boxes and the respective class probabilities for objects whose center is within that cell. Unlike previous methods that performed multiple passes over the image, YOLO processes the entire image at once, significantly contributing to its speed. This simple yet effective architecture allowed YOLO to overcome many of the challenges faced by previous models, including detection in scenarios with objects of different scales and proportions.

The structure of a YOLO model is composed of three main parts: Backbone, responsible for extracting features from the input image; Neck, which connects the Backbone to the Head and is responsible for combining the extracted features at different scales; and Head, which is the part of the model that makes the final predictions. Processes the features combined by the neck and generates the bounding boxes along with the classes of detected objects and the associated probabilities. Each of these parts plays a fundamental role in object detection.

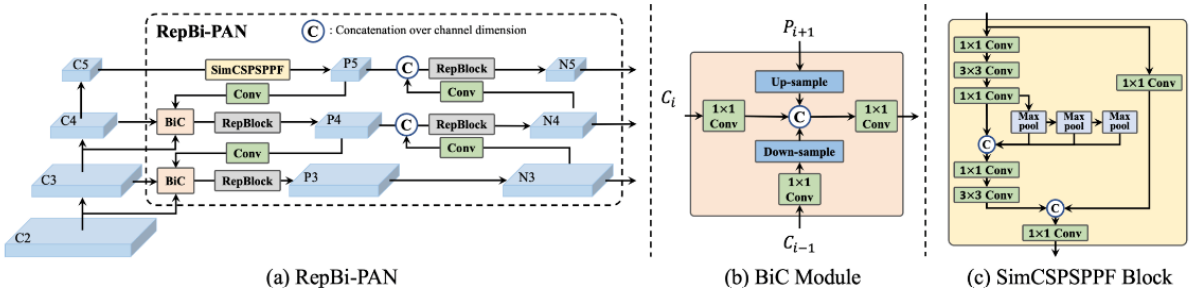
The first version, according to Peiyuan Jiang (J. et al., 2021), had two drawbacks: imprecise positioning and a lower recall rate compared to the region-based recommendation method. It featured a backbone based on the GoogLeNet network (Redmon et al., 2016). The second version of YOLO, known as YOLOv2, developed by Redmon and Farhadi in 2017 (Redmon and A., 2017), brought significant improvements over the original version, mainly in accuracy and the correction of object localization errors. One of the main innovations of YOLOv2 was the introduction of anchors, which are predefined shapes of bounding boxes used to facilitate object detection. These anchors were created based on common shapes for objects of the same class, simplifying the network's learning process. YOLOv2 also introduced a new backbone, Darknet-19, composed of 19 convolutional layers followed by 5 max-pooling layers.

In 2018, Redmon and Farhadi (Redmon and A., 2018) introduced further innovation in YOLO, presenting YOLOv3, which continues to use bounding boxes with anchors. However, in this version, they introduced Multi-Scale Head, which predicts Bounding Boxes and classes at three different scales to detect large, medium, and small images using input image dimensions reduced by 32, 16, and 8 times. They also made a change in the neck to Feature Pyramid Networks (FPN) and switched the backbone to Darknet-53, a deeper network with 53 convolutional layers. All these enhancements contributed to better YOLO performance.

YOLOv4, developed by Alexey Bochkovskiy in 2020 (A. et al., 2020), introduced several innovations that enhanced both the efficiency and accuracy of real-time object detection. Among the main improvements, it featured a real-time object detection system on a GPU and switched the backbone to CSPDarknet53. CSP stands for Cross Stage Partial connections, which has 53 convolutional layers that improve feature extraction capability and reduce computational complexity. Another key innovation was Self-Adversarial Training (SAT), introduced to strengthen the model's robustness. SAT allows the model to self-attack during training, improving its resistance to various types of adversarial attacks, among other enhancements.

YOLOv5, developed by Ultralytics (Ultralytics, 2021), represents an advancement in object detection, maintaining a balance between accuracy and real-time speed. One of the most notable innovations is the adoption of an anchor-free split head, inspired by more recent models like YOLOv8. This approach replaces the traditional predefined anchor boxes with a more flexible mechanism, allowing for superior performance in various scenarios. Additionally, YOLOv5 offers a range of pre-trained models tailored to different needs, ensuring optimized solutions for various applications, such as autonomous vehicles and real-time video analysis.

In 2022, (Li et al., 2022) introduced YOLOv6, which brought advancements such as the EfficientRep Backbone, based on an efficient architecture aimed at improving inference speed and accuracy. YOLOv6 uses Depthwise Separable convolutions and optimization techniques like RepVGG to enhance efficiency. Changes in the Neck include Rep-PAN (RepVGG-PAN), a modified PANet that utilizes RepVGG techniques to improve feature aggregation performance, making the model lighter and faster. Lastly, the Decoupled Head separates bounding box detection and classification into different layers, allowing for more refined optimization for each task, as can be seen in the figure 2.4. In 2023, YOLOv6 v3.0 was introduced in the study titled "YOLOv6 v3.0: A Full-Scale Reloading." (Li et al., 2023) .This model brings significant advancements in architecture and training, incorporating a Bidirectional Concatenation (BiC) module, an Anchor-Assisted Training (AAT) strategy, and an improved backbone and neck design, resulting in high accuracy on the COCO dataset. The following image, as mentioned, better demonstrates the summary of YOLO's body.



(a) RepBi-PAN (b) BiC Module (c) SimCSPSPPF Block

Figure 2.4: Model architecture diagram showing the redesigned network components and training strategies that led to significant performance improvements. (a) The neck of YOLOv6 (N and S are shown). Note that for M/L, RepBlocks is replaced by CSPStackRep. (b) The structure of a BiC module. (c) A SimCSPPPF block. Source (Li et al., 2023)

In 2022, YOLOv7 was introduced (Wang et al., 2022) as a state-of-the-art real-time object detector, featuring significant improvements in speed and accuracy. The model uses the Extended Efficient Layer Aggregation Network (E-ELAN) backbone, which optimizes the network's efficiency and depth through enhancements in layer aggregation to capture features at different depths. In the neck, YOLOv7 continues to use the Path Aggregation Network (PAN) to combine features from different resolutions and incorporates the Extended CSPNet to improve gradient flow and efficiency. Its head, called the Task-Aligned Head, aligns with the features learned by the backbone and neck, enhancing the combination of detection and classification tasks. With a remarkable accuracy of 56.8% Average Precision (AP) and superior performance compared to other detectors like YOLOR and YOLOX, YOLOv7 sets a new standard by being trained from scratch on the MS COCO dataset without relying on pre-trained weights.

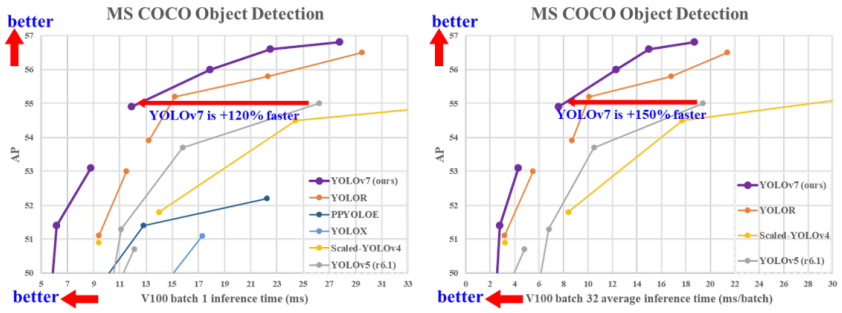


Figure 2.5: YOLOv7 Comparation. Source (Wang et al., 2022)

Other versions of YOLO, such as YOLOv9 and YOLOv10, will not be discussed in this work because they are relatively minor updates to YOLOv8, offering only incremental improvements rather than introducing significant architectural innovations. These versions primarily focus on refining certain aspects, such as optimization techniques and parameter tuning, rather than implementing groundbreaking changes in feature extraction, detection mechanisms, or training strategies. Given their nature as extensions, they do not bring enough novelty to warrant a detailed discussion in this context.

YOLOv8 and YOLOv11, on the other hand, will be discussed in their own chapter, i.e., Chapter 4, due to their more substantial contributions to the evolution of the YOLO family. YOLO proves to be an excellent tool for solving problems involving object detection, whether it is for detecting traffic participants, such as cars, trucks, pedestrians, traffic signs, and lights, regardless of weather conditions, as seen in *The Real-Time Detection of Traffic Participants Using YOLO Algorithm* (Ćorović et al., 2018), or for face detection, as demonstrated in YOLO-face (C. et al., 2020), which is based on YOLOv3 but uses anchor boxes more suitable for face detection. Its adaptability and efficiency make it a popular choice across various domains, enabling robust detection performance even in challenging environments. A structural summary of YOLOS can be seen below 2.1

Version	Backbone	Neck	Head
YOLOv1	Custom CNN (Inspired by GoogLeNet)	No separate "Neck"	Direct prediction using FC layers
YOLOv2	Darknet-19	No separate "Neck"	Anchor boxes introduced
YOLOv3	Darknet-53	No separate "Neck"	Multi-scale detection head
YOLOv4	CSPDarknet-53	Path Aggregation Network (PAN)	Multi-scale detection head
YOLOv5	Cross Stage Partial Networks (CSPNet)	PANet + FPN	Three-layer detection
YOLOv6	EfficientRep (optimized version of CSPNet)	Rep-PAN (optimized PANet)	Multi-scale detection
YOLOv7	Extended Efficient Layer Aggregation Networks (E-ELAN)	PANet	More efficient head
YOLOv8	Custom Backbone (Based on CSPNet)	PANet + FPN	Multi-scale detection, optimized for speed and accuracy

Table 2.1: Comparison of YOLO versions: Backbone, Neck, and Head.

3 RELATED WORKS

This section reviews previously studied works that are directly related to the topic of this research. Relevant studies will be discussed, focusing on similar techniques for photographic curation, image analysis using artificial intelligence, and the development of automated tools for photo selection and enhancement.

3.1 CHOOSE BEST PHOTO

A necessary next step is the selection of the best photo, as we discussed after grouping similar photos. We need to analyze each photo within the subgroup of similar images to determine which one is the best. Choosing the best photo will determine the overall quality of the work and the visual impact of the final result. This task involves not only selecting the best technically suitable image but also considering the emotional aspect that one wishes to convey, which can vary according to individual preferences. Automatic methods for image selection, as proposed in various research studies, employ different approaches to achieve the desired outcome.

A method proposed by (Huang et al., 2022) in “Series Photo Selection via Multi-View Graph Learning” adopts two-branch networks with shared parameters to process different feature views of the same image and compare them. Initially, it employs a Feature Extraction Unit (FEU) to uniformly process multiple feature views of an image, forming multiple views. Subsequently, a graph structure is constructed with these concatenated views, and a two-layer Graph Convolutional Network (GCN) is utilized to simultaneously learn feature information and graph structure. A self-attention module is employed to reinforce common information across different views and maximize consistency. Features are fused and connected to obtain new representations, which pass through a Multilayer Perceptron (MLP) to map the desired function, using cross-entropy loss to compare pairs of images. This method integrates feature and structure learning, promoting automatic and effective selection of the best image.

Another method proposed by (Chang et al., 2016) in “Automatic Triage for a Photo Series” involves a collaborative study using Amazon Mechanical Turk (MTurk), a crowdsourcing platform developed by Amazon. This platform allows companies and researchers to outsource tasks that require human intelligence to a large group of workers distributed worldwide. The task aimed to learn human preferences in photo selection. In this study, participants made pairwise comparisons of photos, choosing their preferred option using a forced-choice methodology and responding to the question, “Imagine you took these two photos and can only keep one. Which would you choose and why?" This approach was used to better measure subtle differences between photos, with participants providing comments justifying their choices. These data were then used to train a model based on the Bradley-Terry algorithm, which assesses the probability of one photo being preferred over another, enabling the creation of a global ranking of photos. This ranking was utilized to train a neural network that automates the selection of the best photo in a series, as shown in the figure 3.1, considering both technical and emotional aspects aligned with the collected human preferences.

In the article (Shah and Kwatra, 2012) ”All Smiles: Automatic Photo Enhancement by Facial Expression Analysis," the implemented methodology involves a sophisticated process to enhance group photographs by leveraging facial expression analysis and pose estimation. The system begins by processing a set of images capturing the same or similar scenes. Initially, it employs off-the-shelf tools to detect faces in each image and extract key facial landmarks,

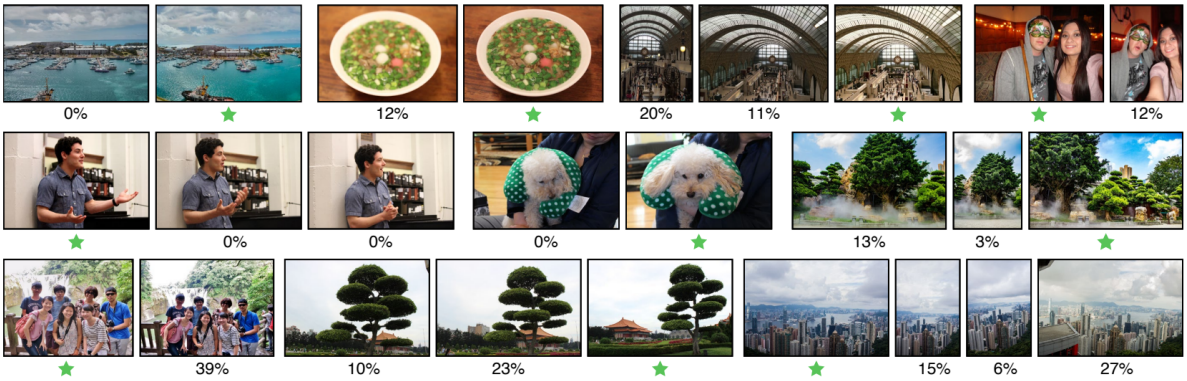


Figure 3.1: The starred photo in each series is the one preferred by the majority of people. Source (Chang et al., 2016) - Automatic Triage for a Photo Series.

including eye centers, the tip and root of the nose, eyebrow corners, and lip corners. These landmarks are essential for performing facial pose identification, where the system determines face orientation in terms of yaw, pitch, and roll, enabling a deeper understanding of each individual's alignment within the frame.

Following this, the system assigns identities to the detected individuals by grouping corresponding faces of the same person across different photos. This is achieved through similarity scores calculated from facial templates, which ensure reliable identification even under variations in pose or expression. The next step involves an automatic evaluation of facial quality using classifiers trained on attributes such as smiling versus non-smiling and open versus closed eyes. These classifiers return continuous scores that reflect the quality of each face, considering both expression and clarity. The scores are then combined with pose information to calculate an overall quality score for each face.

Finally, the system synthesizes a composite image by selecting the input photo with the highest overall quality score. To further enhance the result, it replaces low-scoring faces with high-scoring faces of the same individual from other images, ensuring an optimized composition that highlights the best expressions and poses for all individuals. This meticulous approach enables the creation of a cohesive and visually pleasing group photograph, balancing individual attributes and overall harmony, An example of how the algorithm works is illustrated in the figure 3.2 bellow.

Figure 3.2: Image selection and image after enhancement from article All Smiles: Automatic Photo Enhancement. Source (Shah and Kwatra, 2012)

The choice of the best photo in a set of images is crucial for determining the quality and visual impact of the final result in any automatic photo enhancement system. This process is not limited to merely selecting the image with the best visual technical quality but also involves emotional and aesthetic considerations that vary according to individual preferences. Automatic methods for photo selection have been extensively explored in the literature, each employing distinct approaches to achieve effective results. These advancements are essential to ensure that the photo selection process not only optimizes technical quality but also captures the desired emotional essence in the final visual representation. In this article, we will explore the criterion of open eyes to define the best photos, thereby removing images where eyes are not open from the photographer's selection process.

3.2 FACE DETECTION

Face detection is a widely explored research field and fundamental to various applications in computer vision, from security systems to image curation. One of the most influential approaches is the method proposed by Paul Viola and Michael Jones in 2001 (Viola and Jones, 2001). This method uses cascade classifiers that operate on features extracted using Haar filters, which are mathematical functions used to extract specific characteristics from images, enabling real-time face detection with high accuracy. The key to the success of this method lies in the cascade structure, which applies sequential filters, quickly discarding regions of the image less likely to contain a face, and focusing processing on areas with a higher probability.

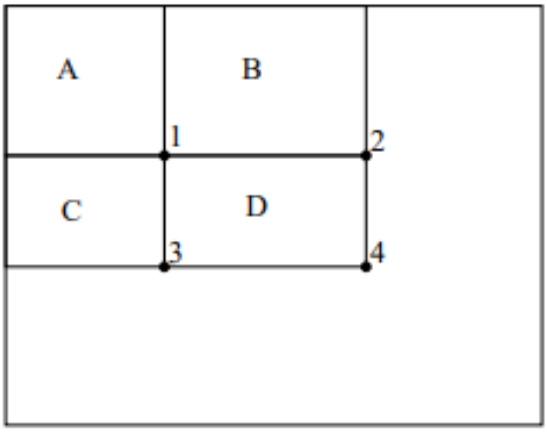


Figure 3.3: The sum of the pixels in a region can be calculated with 4 references. The integral image value at position 1 is the sum of the pixels in rectangle A. The value at position 2 is A + B, at position 3 is A + C, and at position 4 is A + B + C + D. The sum of the pixels in region D can be calculated as 4 + 1 - (2 + 3). Source (Viola and Jones, 2001)

The Adaboost algorithm, introduced by Linhard Maldy in 2002 (Lienhart and Maydt, 2002), also plays a crucial role in this field. It uses the integral image technique, which speeds up the calculation of features, making the detection process more efficient. The combination of Adaboost with the Viola-Jones cascade classifiers results in a robust detection system that can be applied to a variety of lighting conditions and capture angles.

In addition to these classic approaches, modern methods based on convolutional neural networks (CNNs) have revolutionized face detection. These networks are trained on large datasets, learning to identify complex patterns that correspond to faces, regardless of variations in expression, pose, or lighting. The accuracy of these networks allows not only face detection but also the identification of critical facial features, such as eye centers, nose tip, and corners of the

lips. These features are essential for subsequent tasks such as facial recognition and expression analysis.

The use of these modern tools is seen in the article “All Smiles: Automatic Photo Enhancement by Facial Expression Analysis." (Shah and Kwatra, 2012). The implemented methodology involves processing a set of images that capture the same scene or similar scenes. Initially, the system uses ready-made tools to detect faces in each image and extract significant facial landmarks, such as the centers of the eyes, the tip and root of the nose, the corners of the eyebrows and the corners of the lips. Additionally, facial pose identification is performed, determining the face’s orientation in terms of yaw, pitch, and roll, based on the locations of the landmarks. Subsequently, the system assigns identities to the detected people, grouping the corresponding faces of the same person in different photos based on similarity scores derived from facial templates. Another example of using CNNs for face detection is seen in YOLO-face (C. et al., 2020), which is based on YOLOv3. The authors propose a face detector aimed at improving YOLOv3’s performance in face detection. They use the same backbone as YOLOv3 (Darknet-53), a highly efficient and powerful convolutional neural network architecture known for its ability to balance speed and accuracy. However, to adapt the model for face detection tasks, modifications are made to the anchor boxes, which play a critical role in the object detection process. These changes are specifically designed to ensure that the bounding boxes generated by the model are better suited for detecting faces, which often have different aspect ratios and sizes compared to other objects typically detected in general-purpose object detection models. By tailoring the anchors, the model becomes more sensitive to the unique features of faces, such as their relatively small size and specific proportions. This enables YOLO-face to achieve high precision and recall in face detection tasks, making it especially effective in scenarios with challenging conditions, such as varying lighting or occlusions. As illustrated in the figure 3.4, these adjustments significantly enhance the model’s capability to localize and identify faces accurately, building on the robust foundation provided by the original YOLOv3 architecture.

Figure 3.4: Some face detection result examples. Left: results detected by YOLOv2. Middle: results detected by YOLOv3. Right: results detected byYOLO-face. Source (C. et al., 2020)

Face recognition plays a fundamental role in the detection and analysis of the eyes since defining the eyes as a target requires a solid foundation for face detection and subsequent analysis of the eyes of the involved characters.

3.3 EYE DETECTION

After face detection, identifying eyes in images is a crucial step, especially in applications that aim to ensure that people have their eyes open in photos. According to Kim (Kim et al., 2017) in "A study of deep cnn-based classification of open and closed eyes using a visible light camera sensor", methods for classifying whether an eye is open or closed can be categorized into four groups: non-image-based methods, video-based methods, image-based methods without training, and image-based methods with training. In non-image-based methods, the analysis is conducted through brain electrical activity, captured by sensors connected to the ocular muscles. Although these methods are fast, the requirement to connect sensors to the user's body limits their mobility. In contrast, video- or image-based methods do not have this limitation, allowing information to be gathered from images captured from a distance. Eye detection is often performed using template matching techniques, where a specific region of the face is compared with a predefined eye template. The work of Bhoi and Mohanty (Bhoi and Mohanty, 2010) exemplifies this approach, where various portions of the face are compared with an eye template, and the region with the highest match is identified as containing the eye.

As mentioned several times in the article All Smiles (Shah and Kwatra, 2012), after recognizing the face and separating the poses and their positions, the quality of the faces is then automatically evaluated using classifiers trained for attributes such as smiling versus not smiling and eyes open versus closed, as illustrated in the figure 3.5, which return continuous scores that are combined with pose information to determine an overall quality score. Based on these evaluations, the system synthesizes a final composite image by selecting the input photo with the best overall score and replacing low-scoring faces with high-scoring faces of the same person, ensuring an optimized final composition of the group photographs.

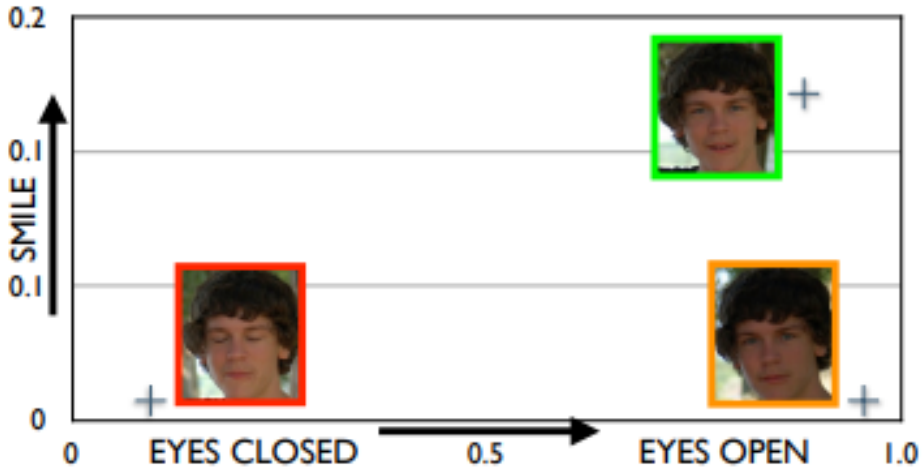


Figure 3.5: Plot showing "open eyes" and "smile" scores for one subject's faces. from article All Smiles: Automatic Photo Enhancement. Source (Shah and Kwatra, 2012)

The use of convolutional neural networks (CNNs) for eye detection has also shown promising results. CNNs can be specifically trained to identify not only the presence of eyes but also their state (open or closed). This is particularly useful in scenarios like image curation, where multiple photos of the same scene are taken to ensure that all individuals have their eyes open. A CNN trained to automatically detect the state of the eyes can significantly reduce the time needed to select the best photos, automating a process that traditionally requires manual inspection. As seen in the article "An Eye Detection Method Based on Convolutional Neural Networks and

Support Vector Machines" (Mingxin et al., 2018), the authors propose an eye detection method that integrates three main techniques: an Eye Verification Filter (EVF), Convolutional Neural Networks (CNNs), and Support Vector Machines (SVMs). The EVF is used for initial filtering, checking whether the detected regions may contain eyes based on predefined patterns. The CNNs, in turn, are specialized neural networks in extracting complex features from images, allowing the model to capture specific details of the eyes. SVMs are classification algorithms that, after feature extraction by the CNNs, are responsible for classifying regions as containing eyes or not. In the testing phase, a cascade face detector locates the facial region, which is then manually adjusted if necessary. The search for eyes is limited to the upper half of the face, where they are most likely to be found. The process starts with the EVF, which filters out irrelevant regions, followed by the CNN, which extracts features from the eye candidates, and ends with the SVM, which performs the final classification, ensuring precise and efficient detection. As illustrated in the figure 3.6

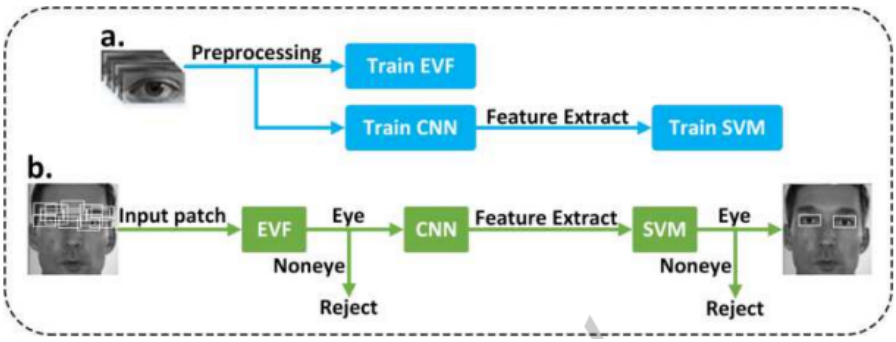


Figure 3.6: (a) Training of eye detection framework and (b) Testing of eye detection framework. Source (Mingxin et al., 2018)

An interesting application of eye detection can be seen in the study “Real-Time Deep Learning-Based Drowsiness Detection: Leveraging Computer-Vision and Eye-Blink Analyses for Enhanced Road Safety" (A.B. et al., 2023). The authors propose an innovative solution to detect drowsy drivers on the road with the aim of preventing accidents. Using deep learning and computer vision techniques, the system focuses on the analysis of eye blinks and yawns, monitoring the driver’s alertness in real time. They implemented a system that detects the driver’s face and, as a first step, identified drowsiness levels based on yawning. Then, the system detects the eyes and calculates blinks, analyzing whether the eyes are open or closed. Additionally, the geometric position of the eyes is also considered as a critical feature in the detection process. This consideration is particularly useful in scenarios where the driver’s head may deviate from standard positions, such as tilting or leaning, which often occurs when the driver begins to fall asleep. By analyzing the relative positions and alignment of the eyes, the system can robustly account for these variations and maintain accurate detection, even under challenging conditions. This geometric analysis not only improves the detection of closed eyes—a common sign of drowsiness—but also enhances the system’s ability to identify subtle shifts in the driver’s gaze or head orientation. This approach ensures reliability in real-world applications, such as driver monitoring systems, where factors like poor lighting, rapid movements, or partial occlusions might otherwise reduce accuracy. As illustrated in the figure, this method integrates spatial and geometric cues to deliver a comprehensive and adaptive monitoring solution. As illustrated in the figure 3.7

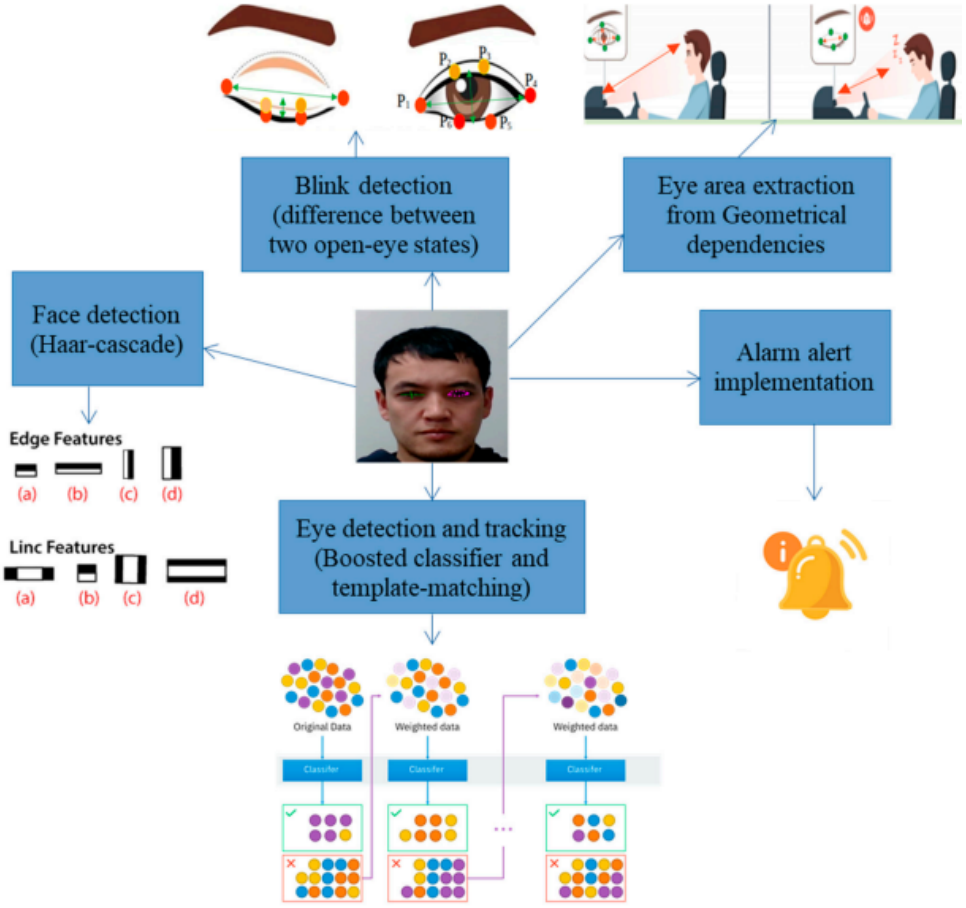


Figure 3.7: Workflow of the proposed method with algorithm application, eye tracking, and distance measurement of eye states and between the camera and driver's head location. Source (A.B. et al., 2023)

The system combines these mechanisms to determine the level of drowsiness. If it detects that the driver is falling asleep or showing signs of fatigue, a sound alert is triggered. The study achieved impressive results, with a 95.8% accuracy rate in detecting drowsy eyes and 97% for open eyes. Furthermore, yawn detection was classified with precision 84%, which reinforced the effectiveness of the proposed method in enhancing road safety.

Another significant application of eye detection, specifically focusing on eye blinking, is proposed in the study "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern" (Jung et al., 2020). This work addresses the growing issue of deepfakes, particularly their use in spreading fake news and damaging reputations. Deepfakes have become increasingly sophisticated and harder to detect, with traditional pixel-based detection methods proving less effective as GANs (Generative Adversarial Networks) have advanced. As illustrated in the figure 3.8

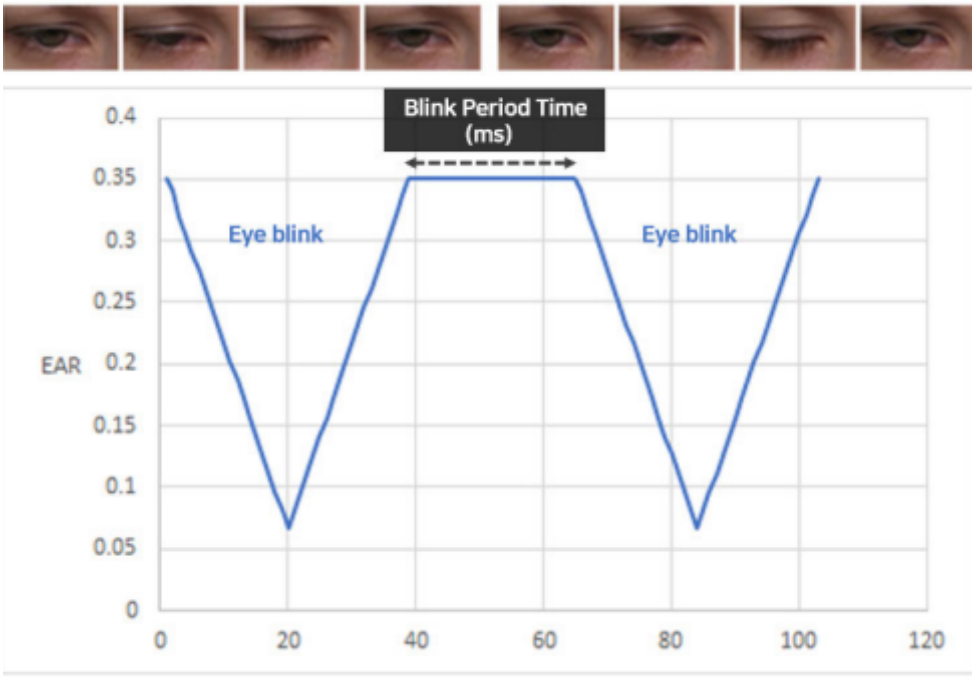


Figure 3.8: This figure shows the method used to measure the eye blink period. Source (Jung et al., 2020)

In this study, the authors developed a method that detects deepfakes by analyzing the natural blinking patterns of human eyes, which deepfakes often fail to replicate accurately. Using the Eye Aspect Ratio (EAR), formula bellows in figure 3.9, a technique originally proposed by Tereza Soukupova and Jan Cech (Soukupov and Cech, 2016), the algorithm identifies six specific points around the eyes to calculate the horizontal and vertical dimensions.

$$\mathrm{EAR} = \frac{||p_2 - p_6|| + ||p_3 - p_5||}{2||p_1 - p_4||}$$

Figure 3.9: EAR formula. Where p1,. . ., p6 are the 2D landmark locations. Source (Soukupov and Cech, 2016)

EAR measures whether eyes are open or closed and tracks blinking behavior. This analysis is based on the fact that human blinking patterns are involuntary and are affected by various cognitive and physiological factors, making them difficult to fake convincingly.

The authors demonstrated the effectiveness of their method, showing that DeepVision could detect anomalies in blinking behavior in seven out of eight deepfake videos tested. They combined this eye tracking with facial detection using the Fast-HyperFace algorithm to improve the detection accuracy. This work presents a promising solution to the challenge of deepfakes and could have broader applications in areas requiring integrity verification, particularly in media and cybersecurity.

In the article “Eye Recognition by YOLO for Inner Canthus Temperature Detection in the Elderly Using a Transfer Learning Approach. " (Ghourabi et al., 2023) the authors use YOLO for eye detection in thermal images. YOLO was used to measure the temperature in the inner corner of the eyes in the elderly, with the goal of identifying early signs of physical frailty and infectious diseases, such as fever, as illustrated in the figure 3.10. The work utilizes thermal images to accurately detect the temperature of the inner canthus, which is the most reliable area for measuring body temperature using infrared cameras. It compares different versions of YOLO (YOLOv5, YOLOv6, YOLOv7), applying a transfer learning approach to a dataset of thermal

images of the elderly. The results indicate that YOLOv7 showed the best performance in terms of accuracy and speed.

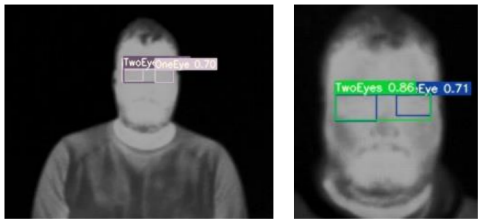


Figure 3.10: Distance effect on eyes detection by YOLOv7 model trained on the augmented multi-age dataset. Source (Ghourabi et al., 2023)

Eye detection is essential for many fields including photo curation as it ensures that all people in an image have their eyes open, significantly improving the quality of the selected photos. In scenarios such as events or group photo sessions, where multiple images of the same scene are captured, this automated detection can speed up the selection process, avoiding the need for manual review. By automatically identifying and prioritizing the best photos, eye detection systems ensure that only high-quality images are chosen, optimizing the final result.

3.4 CONCLUDING REMARKS

This chapter presented several advanced methods and techniques in photo curation, with a specific focus on detecting closed eyes in characters involved in a scene, aiming to assist in selecting the best photos using artificial intelligence and image processing algorithms. We discussed relevant studies that highlight the importance of automated processes to optimize photographers' work, especially in events and photo shoots where multiple images are captured. Therefore, in this work, we will address the development of a tool that helps photographers detect images where eyes are closed, laying the foundation for a broader solution that automates the photo curation process.

4 MATERIALS AND METHODS

This chapter presents the materials and methods used to achieve the objectives proposed in this work. Initially, the construction of a proprietary database is detailed, composed of images from photographic events, which were carefully annotated to identify open and closed eyes. These annotations use two distinct classes: Open-eye and Closed-eye. In the context of computer vision, an annotation class represents a specific category or label assigned to objects or regions of interest within an image, enabling supervised learning models to distinguish between different types of data. For this work, the Open-eye class represents eyes that are visibly open, while the Closed-eye class identifies eyes that are visibly closed. These classes form the foundation of the model's ability to accurately detect and classify eye states.

Next, the architectural characteristics and advantages of YOLOv8 and YOLOv11 versions are described, including their variants and specific configurations. Additionally, the training and evaluation process of the models is explained, detailing the strategies and parameters used to ensure a robust and reliable analysis of the results. The use of these annotation classes, combined with precise labeling, ensures that the models can effectively learn and differentiate between the defined eye states, providing the basis for accurate detection and classification.

4.1 MATERIALS

As the object of study in this work involves detecting the state of eyes in photographs from events, it was necessary to create a proprietary dataset of annotated images relevant to the universe of professional photography. These images comprise a complete event from a gender reveal party and a family photoshoot. The dataset was divided into two subsets: a training set containing 981 images (gender reveal party) and a validation set with 151 images (family photoshoot). These images were manually selected and annotated using the Computer Vision Annotation Tool (CVAT), which offers functionalities for labeling objects in images and videos in various formats, such as bounding boxes, polygons, and points.

Annotations were performed considering the state of the eyes. The criteria followed were that eyes are considered open if the pupil and sclera (the white part surrounding the iris) are visible. Otherwise, the eyes were annotated as closed, example Open-eye class annotation example in the figure 4.1. Additionally, in specific cases where only one eye is visible in the image, the annotation with the class Open-eye was adopted. The impact of this choice on the results will be discussed in Chapter 5, i.e., the Results.

Figure 4.1: Example of an annotation using CVAT. Source (The Author)

The annotations were exported in a format compatible with YOLO, containing the object class identifier (open or closed eye) and the central coordinates (x, y), as well as the width and height of the bounding box, normalized relative to the original image size. This format enables direct integration with detection models.

The annotation process resulted in a robust dataset(Tavares, 2024), which will be used to train and evaluate the detection models developed in this work. The distribution of annotated eyes and the model's performance in specific scenarios, including those with one eye open, will be presented and analyzed in detail in subsequent chapters .

4.1.1 YOLOv8

YOLOv8, one of the most recent versions of YOLO, was released on January 10, 2023, marking a significant milestone in the field of object detection. It introduced architectural advancements that make it more efficient and adaptable to various tasks, especially for detecting small objects. Its modular design, combining improvements in the backbone, neck, and head, enables robust and flexible performance across multiple applications.

Architecture: Backbone, Neck, and Head **Backbone:** The backbone of YOLOv8 is responsible for extracting features from the input image. YOLOv8 employs mechanisms such as CSPNet (Cross Stage Partial Networks) to reduce redundancies in feature extraction, improving computational efficiency and model generalization.

Neck: YOLOv8 uses the neck to combine features extracted at different scales, allowing for more accurate detection of objects of varying sizes. Structures such as the Path Aggregation Network (PAN) are utilized to merge information from both deep and shallow layers, facilitating the detection of both small and large objects.

Head: The head is responsible for making the final predictions, including object classes, bounding box coordinates, and confidence scores. In YOLOv8, the head has been redesigned to be more modular, enabling specific adaptations such as segmentation, pose detection, classification, and other applications.

YOLOv8 is available in four distinct sizes, as follows:

- **YOLOv8-Nano (N):** The smallest and lightest version, designed for devices with extremely limited resources, such as microcontrollers, IoT, and basic embedded systems. Ideal for real-time applications requiring high energy efficiency and low latency, even on hardware without a GPU.

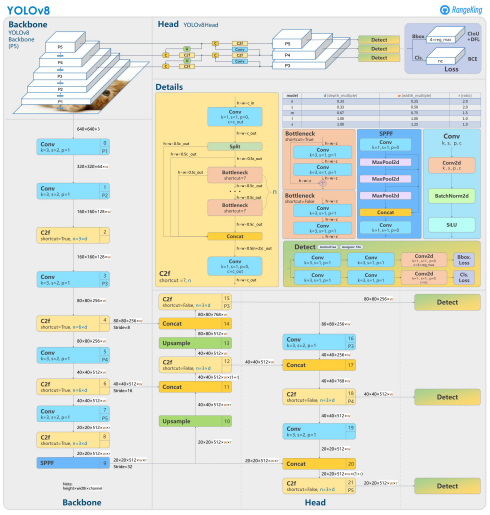


Figure 4.2: YOLOv8 architecture. Source (Jocher et al., 2023)

- **YOLOv8-Small (S):** A lightweight and compact model suitable for devices like smartphones and more robust embedded systems. It offers a good balance between performance and resource consumption in standard object detection scenarios.
- **YOLOv8-Medium (M):** Represents a more advanced balance between accuracy and efficiency, making it ideal for general applications that require good performance with moderate computational costs.
- **YOLOv8-Large (L):** Focused on higher accuracy, this model is more suitable for systems with greater computational capacity, such as servers or high-performance GPU devices.
- **YOLOv8-Extra Large (X):** The most robust and powerful version, focused on delivering the highest possible accuracy in complex and demanding scenarios. It requires significant computational resources, making it ideal for cutting-edge applications and use in data centers.

In addition to standard object detection, YOLOv8 has been extended to cater to different scenarios and specific tasks, featuring the following:

- **YOLOv8-SEG:** Designed for instance segmentation tasks, where it is necessary to identify not only the bounding boxes of objects but also their precise contours in a binary or multi-class mask.
- **YOLOv8-POSE:** Specialized in human pose detection, predicting keypoints to map body joints and positions. This version is widely used in sports monitoring, motion analysis, and augmented reality.
- **YOLOv8-OBB (Oriented Bounding Boxes):** Introduces oriented bounding boxes for detection in images with tilted or irregularly oriented objects, such as aerial photos or industrial inspections.
- **YOLOv8-CLS:** A model tailored for image classification tasks, enabling the categorization of an entire image into one or more classes.

Figure 4.3: Ships detection using YOLOv8-OBB. Source (Jocher et al., 2023)

YOLOv8 offers a modern and modular architecture that combines efficiency, flexibility, and precision. Its variants allow the model to be adapted to a wide range of applications, from embedded devices to robust servers. Specific versions, such as YOLOv8-SEG and YOLOv8-POSE, expand its usability, making it a versatile choice for complex computer vision problems. The results obtained from the implementation and evaluation of the different variants and versions of YOLOv8 will be discussed in detail in the Results chapter, where we will analyze its performance in relation to the objectives of this work.

4.1.2 YOLOv11

YOLOv11 is the latest evolution of the YOLO series, launched in October 2024, bringing significant improvements in efficiency and performance for computer vision tasks. It was designed with a focus on optimizations for more complex tasks while maintaining the philosophy of being a fast and modular model. YOLOv11 builds upon the foundation established by YOLOv8, incorporating innovations introduced in YOLOv9 and YOLOv10, solidifying itself as the most advanced and versatile version in the YOLO family.

4.1.2.1 Architecture: Backbone, Neck, and Head

Backbone YOLOv11 introduces changes such as:

- **Conv Layers:** Similar to YOLOv8, it starts with convolutions to reduce the image resolution.
- **C3k2 Block:** Replaces the C2f block used in previous versions, introducing smaller and faster convolutions, optimizing feature extraction.
- **CSP (Cross Stage Partial):** Implemented to split and process feature maps, reducing computational load and improving data representation.
- **C2PSA Block (Cross Stage Partial with Spatial Attention):** Added after the SPPF block, this mechanism enhances spatial attention, allowing the model to focus on important regions of the image with greater precision.

Neck Key changes include:

- **C3k2 Block:** Replaces the C2f in the neck as well, improving information aggregation and increasing efficiency.
- **Spatial Attention Mechanism:** Incorporated through the C2PSA block, ensuring better focus on small or partially occluded objects.

Head The head includes the following changes:

- **C3k2 Block:** Continues the replacement of C2f to optimize inference.
- **Detect Layer:** Preserves YOLOv8's schema, adjusting predictions at different resolutions.

Like YOLOv8, YOLOv11 is available in different sizes (**Nano, Small, Medium, Large, and Extra Large**), adapting to various needs for precision and computational capacity. Additionally, it also supports specialized variants like **SEG, POSE, OBB, and CLS**, offering specific functionalities for segmentation, pose detection, oriented bounding boxes, and image classification.

YOLOv11 consolidates architectural advancements such as the introduction of the C3k2 and C2PSA blocks, ensuring greater efficiency, precision, and versatility. It represents a milestone in the field of computer vision, capable of serving both embedded devices and cutting-edge applications on high-performance servers. Additional details and performance comparisons will be presented in the following chapters.

4.1.3 Methods

The method used in this work involved training models to learn how to detect the state of the eyes in photographic event images, distinguishing between open and closed eyes. For this purpose, we used previously cataloged images, divided into two subsets: training, with 981 images from a gender reveal event, and validation, with 151 images from a family photo session. This separation was essential to evaluate the models' performance in real-world professional photography scenarios, ensuring a consistent analysis of the results.

4.1.3.1 Training Configuration

The models were trained using YOLOv8 and YOLOv11 in their Medium (M) and Medium-OBB (Oriented Bounding Boxes) versions, configured with the following parameters:

- **Number of epochs:** 500;
- **Batch size:** 16;
- **Image dimension:** 640 pixels (automatically resized by YOLO during preprocessing).

The training was conducted using the standard command provided by the respective YOLO implementations, which includes automatic optimizations to avoid redundancies and accelerate the convergence process. The experiments were carried out on a computer with a Quad-Core AMD Opteron™ 8387 2.8GHz processor, 64GB of RAM, and an NVIDIA Titan X GPU with 12GB of RAM and 3,072 CUDA cores.

4.1.3.2 Training Strategy

Both YOLOv8 and YOLOv11 have built-in mechanisms to automatically stop training if performance stagnates or shows signs of overfitting, based on metrics such as loss and mAP (mean Average Precision). This mechanism significantly reduced the number of effectively used iterations, completing training in less than 250 epochs for both versions.

4.1.3.3 Model Evaluation

After training, the models were evaluated using the validation set. This stage included both quantitative analysis (performance metrics such as precision, recall, and F1-score) and qualitative analysis, examining specific predictions, such as detecting partially visible eyes or handling challenging lighting conditions. Detailed results for each configuration, as well as a comparison between the different YOLO versions, will be presented in Chapter 5.

The approach described in this chapter laid a solid foundation for training and evaluating eye detection models in realistic professional photography scenarios. The annotation, training, and evaluation steps were carefully detailed to ensure the reproducibility of the research.

5 RESULTS

This chapter presents and analyzes the results obtained from the experiments conducted throughout this work. The tests were carried out using different configurations and variants of the YOLOv8 and YOLOv11 detection models, focusing on the task of identifying open- and closed-eyes in event photographs. The analysis aims not only to identify the best-performing model but also to understand how the configurations influence aspects such as reliability, efficiency, and suitability for the specific demands of photographic curation.

We introduce the metrics used for the *Closed-Eye* and *Open-Eye* classification. After, we detail the results, highlighting the main observations and interpretations.

5.1 METRICS

5.1.1 Precision (P)

Precision measures the proportion of correct predictions among all predictions made for a specific class. Practically, it is a metric that evaluates the model's reliability in correctly identifying objects of interest and is computed as follows

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

where:

- **TP (True Positives):** number of correct detections;
- **FP (False Positives):** number of incorrect detections.

A high precision value indicates that the model has few false positives, meaning it rarely classifies something incorrectly as part of the class of interest.

5.1.2 Recall (R)

Recall measures the proportion of relevant objects correctly detected in relation to the total number of actual objects in that class. It evaluates the model's ability to find all objects of interest and is computed as follows

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where:

- **TP (True Positives):** number of correct detections;
- **FN (False Negatives):** number of relevant objects not detected.

A high recall value indicates that the model detects the most relevant objects, even if it sometimes includes incorrect detections.

5.1.3 Mean Average Precision (mAP)

Mean Average Precision is a more comprehensive metric used to evaluate the model's overall performance across different confidence levels and classes. It considers the relationship between precision and recall for varying overlap thresholds (*IoU*, or Intersection over Union).

5.1.4 mAP50

mAP50 is the average precision calculated for a fixed IoU threshold of 50%. This means that a prediction is considered correct if the intersection of the predicted box with the actual box is at least 50% of the total size of their union and it is computed as follows

$$\text{mAP50} = \frac{\sum_{i=1}^{n} AP_i}{n},$$

where:

- AP_i: area under the precision-recall curve for class i;
- n: total number of classes.

High mAP50 values indicate that the model can correctly detect objects with sufficient overlap, even in less strict scenarios.

5.1.5 mAP50-95

mAP50-95 is a stricter version of mAP, as it calculates the average precision for multiple IoU thresholds, ranging from 50% to 95% in increments of 5%. This metric evaluates the model's performance more thoroughly and rigorously, considering scenarios where the predicted boxes need to be almost perfectly aligned with the actual ones. A high *mAP50-95* value indicates that the model is robust and reliable, even in scenarios where predictions need to be very precise.

5.2 EVALUATED CONFIGURATIONS

The YOLOv8 and YOLOv11 versions were tested, including models with standard medium detection and the medium version with Oriented Bounding Box (OBB). These variants were selected for their specific capabilities in detection tasks, such as precision in oriented bounding boxes and efficiency in scenarios with small or complex objects. The use of OBB enhances the detection performance by providing a better fit to objects with varying orientations. This capability is particularly useful in applications where object rotation plays a significant role, such as in detecting rotated objects or in environments with less predictable object alignments.

5.3 OBTAINED RESULTS

The overall performance results were obtained from two distinct data collection and annotation approaches. The first approach used annotations described in Section 4, generated in CVAT, covering different angles and eye positions, and including annotations in photos where only one eye (from two) was visible. This allowed for a broader analysis of the ability of the YOLO models to handle significant image variations. In contrast, the second approach adjusted the annotations to be more binary, focusing exclusively on cases where both eyeballs (iris+pupila) were visible in nearly frontal photos. This adjustment aimed to create a more uniform and simplified dataset that reflects more specific and less ambiguous conditions to evaluate how the models perform in more constrained scenarios.

In the following, we present the results of each of these approaches.

In the first approach, shown in Table 5.1 and Figure 5.1, the original annotations generated in CVAT were maintained, which included various angles, eye positions, and various conditions. This dataset represented a more challenging scenario that required YOLO models

Version	Size	Detection Type	Class	Images	Instances	Precision (P)	Recall (R)	mAP50	mAP50-95
YOLOv11	m	Detection	All	151	391	0.707	0.689	0.691	0.327
	m	Detection	Closed-Eye	56	66	0.493	0.606	0.516	0.247
	m	Detection	Open-Eye	140	325	0.921	0.772	0.867	0.408
YOLOv8	m	Detection	All	151	391	0.737	0.608	0.702	0.371
	m	Detection	Closed-Eye	56	66	0.644	0.411	0.550	0.298
	m	Detection	Open-Eye	140	325	0.830	0.806	0.853	0.443
YOLOv8	m	OBB Detection	All	151	391	0.751	0.653	0.720	0.428
	m	OBB Detection	Closed-Eye	56	66	0.678	0.485	0.576	0.320
	m	OBB Detection	Open-Eye	140	325	0.824	0.822	0.863	0.536
YOLOv11	m	OBB Detection	All	151	391	0.805	0.705	0.768	0.466
	m	OBB Detection	Closed-Eye	56	66	0.669	0.606	0.628	0.341
	m	OBB Detection	Open-Eye	140	325	0.940	0.803	0.908	0.590

Table 5.1: Results of Precision, Recall, mAP50, and mAP50-95 for each version of YOLO configuration for the first approach. Source(The Author)

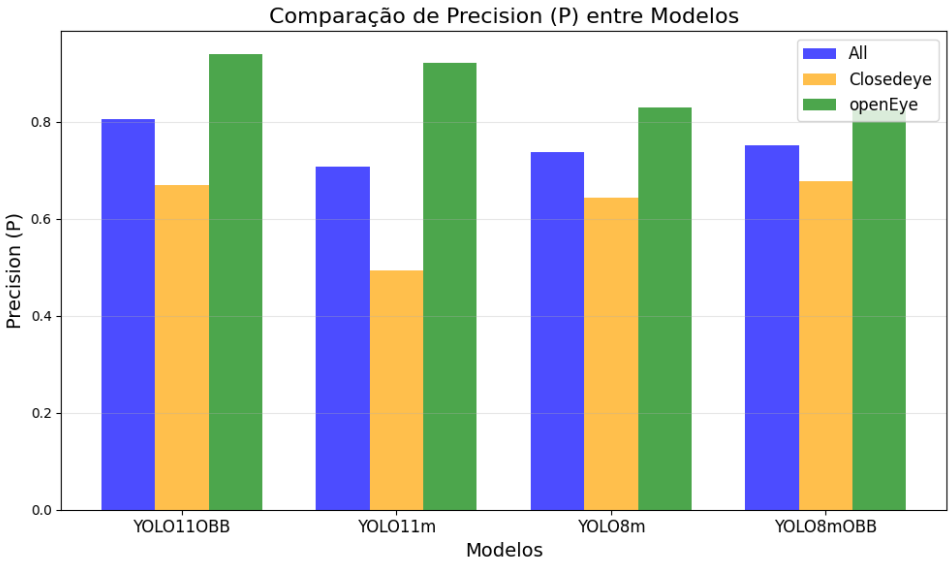


Figure 5.1: Graph with the results of the first approach comparing the accuracy of each version of YOLO. Source(The Author)

to handle greater image variability. YOLO versions that use Oriented Bounding Boxes (OBB) demonstrated superior results compared to standard versions due to their ability to better handle objects oriented at different angles. This feature is particularly important in eye detection tasks, where eye position and alignment can vary significantly between images.

The results showed that the OBB variants improved both precision (P) and recall (R) in all the scenarios evaluated. For example, the YOLOv11-m OBB achieved an overall precision of 0.805, compared to 0.707 for the standard YOLOv11-m version. Furthermore, the recall for the OBB version was 0.705, while the standard version scored 0.691. This difference highlights the OBBs' ability to detect objects more reliably, reducing false positives and increasing true positives.

An even more evident improvement was observed in mean average precision (mAP) metrics. The YOLOv11-m OBB reached an mAP50 of 0.768 and an mAP50-95 of 0.466, surpassing the standard version's values of 0.691 and 0.516, respectively. These results indicate that OBBs provide more accurate alignment between predicted and actual bounding boxes, particularly in more rigorous scenarios where a higher minimum overlap is required (such as in mAP50-95).

Similarly, the YOLOv8-m OBB also outperformed its standard counterpart. Precision increased from 0.737 to 0.751, while recall rose from 0.608 to 0.653. There were also improvements in mAP50 and mAP50-95, with values increasing from 0.702 and 0.371 to 0.72 and 0.428, respectively.

These results confirm that using oriented bounding boxes is advantageous in tasks that demand high accuracy in object alignment. This characteristic makes OBB versions more suitable for applications involving specific details, such as eye detection in photographs, where minor deviations can significantly impact detection quality. Thus, OBBs stand out as an ideal choice for scenarios requiring robustness and reliability in predictions.

The YOLOv11-m OBB version demonstrated the best overall performance in terms of precision (P) and mAP50-95, especially for the Open-Eye class, achieving an mAP50 of 0.898 and an mAP50-95 of 0.597. This version also showed improvements for the Closed-Eye class compared to the YOLOv8-m OBB. Based on these results, the YOLOv11-m version with OBB detection proved to be the most efficient for detecting open and closed eyes, indicating that the adaptation with oriented bounding boxes can be advantageous for tasks requiring precision in angled detections.

Version	Size	Detection Type	Class	Images	Instances	Precision (P)	Recall (R)	mAP50	mAP50-95
YOLOv11	m	Detection	All	151	391	0.764	0.624	0.686	0.341
	m	Detection	Closed-Eye	56	66	0.658	0.439	0.496	0.253
	m	Detection	Open-Eye	140	325	0.869	0.809	0.876	0.429
YOLOv8	m	Detection	All	151	391	0.802	0.624	0.702	0.359
	m	Detection	Closed-Eye	56	66	0.709	0.455	0.526	0.261
	m	Detection	Open-Eye	140	325	0.895	0.794	0.878	0.457
YOLOv8	m	OBB Detection	All	151	391	0.723	0.705	0.747	0.431
	m	OBB Detection	Closed-Eye	56	66	0.637	0.545	0.625	0.335
	m	OBB Detection	Open-Eye	140	325	0.809	0.865	0.869	0.527
YOLOv11	m	OBB Detection	All	151	391	0.788	0.641	0.726	0.447
	m	OBB Detection	Closed-Eye	56	66	0.685	0.424	0.555	0.319
	m	OBB Detection	Open-Eye	140	325	0.892	0.858	0.896	0.574

Table 5.2: Results of Precision, Recall, mAP50, and mAP50-95 for each YOLO configuration for the second approach.

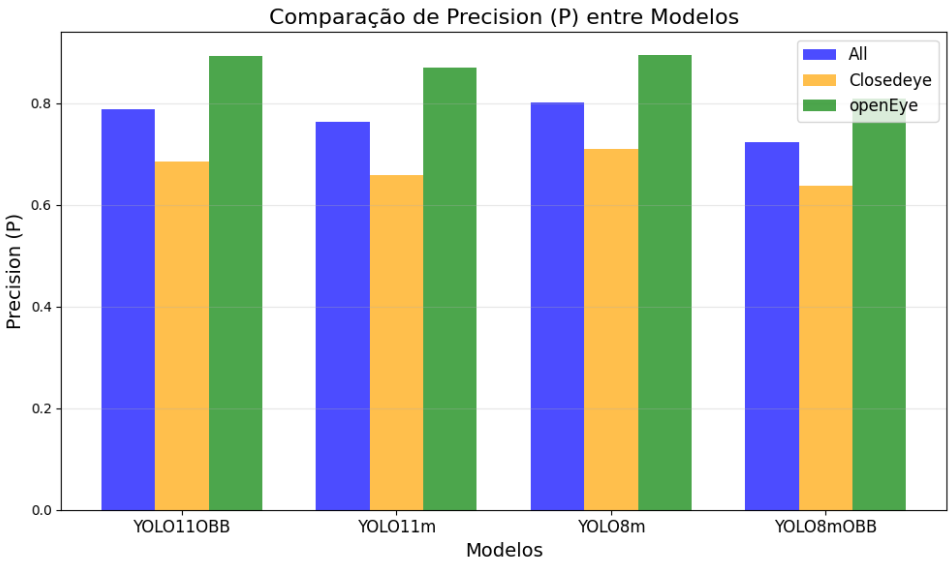


Figure 5.2: Graph with the results of the second approach comparing the accuracy of each version of YOLO. Source(The Author)

In contrast, the second approach, shown in Table 5.2 and Figure 5.2, involved adjusting the annotations to more binary cases, restricting the dataset to nearly frontal images with both eyeballs clearly visible. This less ambiguous dataset provided a simplified scenario, favoring models that rely on greater uniformity in visual features.

In this context, YOLOv8-m stood out as the best overall model, achieving a general precision (P) of 0.802, the highest among all configurations. Although the YOLOv11-m OBB's mAP50-95 remained superior (0.466 in the first approach and 0.597 in the second), the high precision of YOLOv8-m reflects its good balance between simplicity and robustness, particularly in a less challenging scenario.

The second approach demonstrated that standard models, such as YOLOv8-m, can achieve higher general precision when the data is more consistent and simplified, as they are less dependent on advanced techniques like OBBs.

5.4 INTERPRETATION OF RESULTS

The analysis of the results highlighted the impact of the characteristics of the dataset on the performance of detection models, with notable differences between the approaches used. The greater diversity in the first approach, which reflects more realistic scenarios of uncontrolled events, benefited models with Oriented Bounding Boxes (OBB) due to their superior ability to adapt to different angles and orientations. This feature is crucial in detection tasks where object position variability is high, such as in the case of eyes, whose positions can vary significantly due to factors like head rotation or facial expression.

However, the simplification of the data in the second approach resulted in fewer challenges for standard models. This allowed these models to approach or, in some cases, surpass the OBB models in metrics such as overall precision. The reduction in image complexity, achieved by restricting the dataset to more uniform angles and clearly visible eyeballs, decreased the need for advanced techniques such as OBBs. This demonstrates the direct relationship between the choice of the ideal model, the specific characteristics of the dataset, and the application requirements.

OBB models showed clear superiority in more diverse and challenging scenarios (first approach), excelling in situations where object angles and orientations are variable, and standard models performed competitively in uniform and less ambiguous scenarios (second approach), benefiting from their lower reliance on specialized techniques and consequently lower computational cost.

These results demonstrate that, while OBB models are better suited for high-variability scenarios, standard models prove efficient in more controlled applications with less data variability.

Another interesting factor is that open-eye detection showed significantly higher precision compared to closed-eye detection, mainly due to the ease with which the network can identify clear features, such as the iris and sclera. For example, in the YOLOv11-m OBB version, the Open-Eye class achieved a precision of 0.940, reflecting the network's ability to clearly identify the eyes. These elements are easier to distinguish and more consistent across images, which facilitates the neural network's task, especially in well-lit scenarios. In contrast, detecting closed eyes is more subjective and challenging, as it depends on factors such as lighting, eyelid position, and possible reflections, which can create considerable variations in the images. In the same version, the Closed-Eye class achieved a precision of 0.669, highlighting how these factors make the task more difficult. Furthermore, the presence of shadows and reflections around the eyes can further complicate the differentiation between closed eyes and other elements, leading to a higher

error rate and greater sensitivity to false positives. This factor contributes to the performance difference observed between the two detection classes.

Despite the promising performance achieved with the use of oriented bounding boxes, some challenges were identified during the analysis. One of the main issues was the occurrence of false positives, where the model detected objects that were not eyes or incorrectly classified image elements as belonging to the Open-Eye or Closed-Eye classes. These errors can be attributed to the complexity of the images, especially in scenarios with textures or patterns resembling eyes. For instance, shadows, glares, or textured areas around the eyes can confuse the models and lead to erroneous detections.

Figure 5.3: False positive x Correct predict. Source (the author).

Another relevant point was the difficulty of the tested versions in detecting eyes positioned vertically or at extreme angles. In situations where the eye is significantly tilted, the detection failed or presented inconsistencies, which negatively impacted the recall under such conditions. In the validation base, photos 105 to 128 were not detected because they were vertical. This limitation is particularly evident in photographic events with a large diversity of positions and orientations, such as images where the angles differ from the usual.

Figure 5.4: Image with vertical eyes without detection. Source (the author).

The detailed analysis revealed that the use of oriented bounding boxes in YOLOv11 and YOLOv8 significantly contributed to improving precision, especially in situations where the object's orientation plays a critical role. However, YOLOv8 proved to be an efficient alternative for applications that require higher inference speed, although it shows a slight loss in precision compared to YOLOv11.

A deeper analysis revealed that the use of oriented bounding boxes in the YOLOv11 and YOLOv8 versions significantly contributed to improving accuracy. This advantage was particularly evident in situations where object orientation plays a critical role, such as detecting tilted or partially obscured eyes. Meanwhile, YOLOv8 stood out as an efficient alternative for

Figure 5.5: Correct Detection. Source (the author).

applications requiring faster inference speeds, although it exhibited a slight loss in accuracy compared to YOLOv11.

These results indicate that, for applications that require robustness in challenging detections, OBB models are an ideal choice. In contrast, in more controlled scenarios or those with real-time performance constraints, YOLOv8 emerges as an efficient alternative that balances accuracy and speed. This distinction underscores the importance of aligning the model characteristics with the specific requirements of the application.

6 CONCLUSION

This study successfully demonstrated the applicability of state-of-the-art object detection models in automating the identification of closed eyes in event photographs. Through the implementation of two distinct approaches, the research provided a comprehensive analysis of the strengths and limitations of models such as YOLOv8 and YOLOv11, including their OBB (Oriented Bounding Box) variants. The first approach, designed to handle diverse and challenging scenarios, highlighted the superiority of OBB-based models, particularly in detecting objects with varying orientations. The second approach, characterized by a simplified dataset, demonstrated the efficiency of standard models in controlled environments, emphasizing the adaptability of detection strategies to specific application needs.

The results obtained were extremely promising, especially regarding the detection of open eyes, which showed exceptional performance. Both YOLOv8 and YOLOv11 OBB versions excelled by presenting high precision in the OpenEye class, with YOLOv11-m OBB achieving an overall precision of 0.940 and a recall of 0.803. These results indicate the remarkable ability of OBB models to handle variations in eye angles and alignments, providing more accurate detections. The improvement in mAP50 values also reflects the superiority of these models, highlighting that OBBs are particularly effective in more challenging scenarios with greater image variability.

OBB-based models proved to be ideal for dynamic and complex scenarios where object orientation can vary significantly. Meanwhile, standard models, such as YOLOv8-m, showed competitive performance in more controlled and simplified environments, effectively balancing precision and computational efficiency. This analysis reinforces the importance of aligning the model selection with the specific requirements of each task, ensuring optimized results in both challenging and controlled contexts.

In addition to the technical achievements, this work significantly contributes to the evolution of photo curation in professional photography. By automating the detection of open- and close-eye, the proposed solutions not only reduce the time and effort required for manual inspection but also improve the precision and efficiency of the photo curation process. With these results, it is clear that the application of trained networks can transform the way photographers select and edit their images, enabling a more agile and accurate curation process.

Future work can further expand this research by using larger datasets to improve model precision and exploring new features, such as smile detection, pose estimation, and advanced segmentation techniques. Moreover, integrating these detection capabilities into real-time workflows or cloud-based platforms could provide photographers with scalable and accessible tools to significantly improve their productivity and work quality. This study represents an important step forward in the use of artificial intelligence for creative tasks, fostering innovation in both technology and the art of photography.

REFERENCES

A., B., W., C.-Y., and L., H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *CoRR, abs/2004.10934*.

A.B., S. F. A. F. A., R., N., and Y.I., C. (2023). Real-time deep learning-based drowsiness detection: Leveraging computer-vision and eye-blink analyses for enhanced road safety. *Sensors*.

Bhoi, N. and Mohanty, M. (2010). Template matching based eye detection in facial image. *International Journal of Computer Applications,*.

C., W., H., H., P., S., Z., C., and Z., C. (2020). Yolo-face: a real-time face detector. *Springer-Verlag GmbH Germany,*, page 805–8013. Reserch Article.

CBN (2023). Mercado de casamentos reaquece no pós-pandemia e movimenta cerca de r40$biporano$.

Chang, H., Yu, F., Wang, J., Ashley, D., and Finkelstein, A. (2016). Automatic triage for a photo series. *ACM Transactions on Graphics (TOG)*, 32(4):1–10. Reserch Article.

Dettmers, T. (2015). Understanding convolution in deep learning. https://timdettmers.com/2015/03/26/convolution-deep-learning/. Acessado em 19/08/2024.

G. Kovasznay, L. S. and Joseph, H. M. (1955). Image processing. *Proceedings of the IRE*, 43(5):560–570.

Ghourabi, M., Mourad-Chehade, F., and Chkeir, A. (2023). Eye recognition by yolo for inner canthus temperature detection in the elderly using a transfer learning approach. *Sensors*, 23(4).

Huang, J., Zhang, L., Gong, Y., Zhang, J., Nie, X., and Yin, Y. (2022). Series photo selection via multi-view graph learning. *2022 IEEE International Conference on Multimedia and Expo (ICME)*. Reserch Article.

Intelligence, M. (2024). Tamanho do mercado de câmeras digitais e análise de participação – tendências e previsões de crescimento (2024 – 2029).

J., P., E., D., L., F., C., Y., and M., B. (2021). A review of yolo algorithm developments. *The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)*. Reserch Article.

Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics yolov8.

Jung, T., Kim, S., and Kim, K. (2020). Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154.

Kim, K. W., Hong, H. G., Nam, G. P., and Park, K. R. (2017). A study of deep cnn-based classification of open and closed eyes using a visible light camera sensor. *Sensors*, 17(1).

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.

Li, C., Li, L., Geng, Y., Jiang, H., Cheng, M., Zhang, B., Ke, Z., Xu, X., and Chu, X. (2023). Yolov6 v3.0: A full-scale reloading.

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., Li, Y., Zhang, B., Liang, Y., Zhou, L., Xu, X., Chu, X., Wei, X., and Wei, X. (2022). Yolov6: A single-stage object detection framework for industrial applications.

Lienhart, R. and Maydt, J. (2002). An extended set of haar-like features for rapid object detection. *Proceedings of the International Conference on Image Processing*, 1:1–900. Reserch Article.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2.

Mingxin, Y., Xiaoying, T., Yingzi, L., Schmidtb, D., Xiangzhou, W., Yikang, G., and Bo, L. (2018). An eye detection method based on convolutional neural networks and support vector machines. *Intelligent Data Analysis*, pages 345–362.

Redmon, J. and A., F. (2017). Yolo9000: Better, faster, stronger. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6517–6525. Reserch Article.

Redmon, J. and A., F. (2018). Yolov3: An incremental improvement. arxiv. *Computer Vision and Pattern Recognition (cs.CV)*.

Redmon, J., Divvala, S., Girshick, R., and A., F. (2016). You only look once: Unified, real-time object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788. Reserch Article.

Repsly (2024). Average number of photos taken per day around the world. `https://www.repsly.com/blog/field-team-management/field-data-insight-average-number-of-photos-taken-per-day-worldwide`. Acessado em 09/05/2024.

Shah, R. and Kwatra, V. (2012). All smiles: automatic photo enhancement by facial expression analysis. *CVMP '12: Proceedings of the 9th European Conference on Visual Media Production*, pages 1–10. Reserch Article.

Soukupov, T. and Cech, J. (2016). Real-time eye blink detection using facial landmarks.

T. Peli, D. M. (1982). A study of edge detection algorithms. *Computer Graphics and Image Processing*, 20(1):1–21.

Tavares, T. (2024). Dontblink dataset.

Tejani, S. (2016). Machines that can see: Convolutional neural networks. https://shafeentejani.github.io/2016-12-20/convolutional-neural-nets/. Acessado em 19/08/2024.

Ultralytics (2021). YOLOv5: A state-of-the-art real-time object detection system. `https://docs.ultralytics.com`. acessado em 20/08/2024.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognitionn. CVPR 2001*, 1:1–1. Reserch Article.

Voinov, S. (2020). . *Deep Learning-based Vessel Detection from Very High and Medium Resolution Optical Satellite Images as Component of Maritime Surveillance Systems*. PhD thesis, Universität Rostock., Rostock - Germany. 90 pgs.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.

Ćorović, A., Ilić, V., Đurić, S., Marijan, M., and Pavković, B. (2018). The real-time detection of traffic participants using yolo algorithm. *26th Telecommunications forum TELFOR 2018*.