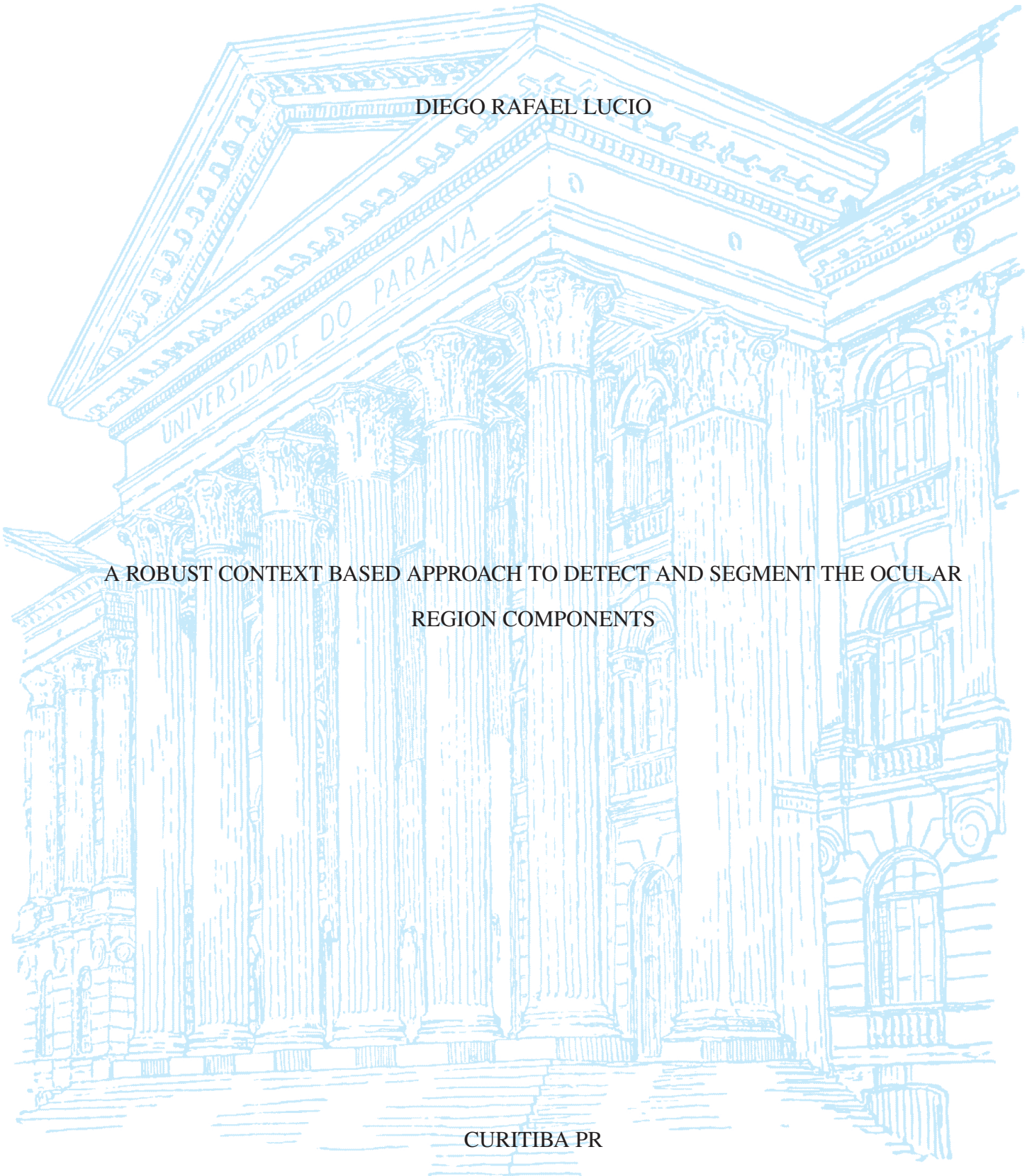
UNIVERSIDADE FEDERAL DO PARANÁ

DIEGO RAFAEL LUCIO



A ROBUST CONTEXT BASED APPROACH TO DETECT AND SEGMENT THE OCULAR REGION COMPONENTS

CURITIBA PR

2022

DIEGO RAFAEL LUCIO

A ROBUST CONTEXT BASED APPROACH TO DETECT AND SEGMENT THE OCULAR REGION COMPONENTS

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Ciência da Computação no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná..

Área de concentração: Ciência da Computação.

Orientador: David Menotti.

Coorientador: Yandre Maldonado e Gomes da Costa.

CURITIBA PR

2022

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

Lucio, Diego Rafael
A robust context based approach to detect and segment the ocular region components / Diego Rafael Lucio. – Curitiba, 2022.
1 recurso on-line : PDF.

Tese (Doutorado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática.

Orientador: David Menotti Gomes
Coorientador: Yandre Maldonado e Gomes da Costa

1. Biometria. 2. Identificação biométrica. 3. Íris (Olhos). 4. Inteligência artificial . I. Universidade Federal do Paraná. II. Programa de Pós-Graduação em Informática. III. Gomes, David Menotti. IV. Costa, Yandre Maldonado e Gomes da. V. Título.

Bibliotecário: Elias Barbosa da Silva CRB-9/1894

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **DIEGO RAFAEL LUCIO** intitulada: **A robust context based approach to detect and segment the ocular region components**, sob orientação do Prof. Dr. DAVID MENOTTI GOMES, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 13 de Maio de 2022.

Assinatura Eletrônica
18/08/2022 10:00:22.0
DAVID MENOTTI GOMES
Presidente da Banca Examinadora

Assinatura Eletrônica
18/08/2022 14:22:09.0
ALCEU DE SOUZA BRITTO JR
Avaliador Externo (PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ)

Assinatura Eletrônica
26/08/2022 14:27:45.0
EDUARDO TODT
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
18/08/2022 11:10:50.0
DIEGO BERTOLINI GONCALVES
Avaliador Externo (UNIVERSIDADE TECNOLOGIA FEDERAL DO PARANA - UTFPR/CAMPO MOURAO)

Assinatura Eletrônica
18/08/2022 09:55:26.0
RODRIGO MINETTO
Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ)

To my parents Jazon Lucio Sobrinho and Ana Cristina Malaguti Lucio, to my brother Felipe Matheus Lucio, and to love of my life Jessica Luana dos Santos.

ACKNOWLEDGEMENTS

First, I would like to thank my parents Jazon Lucio Sobrinho and, Ana Cristina Malaguti Lucio, and my Brother Felipe Matheus Lucio, for always being there for me and for all the strength they have always given me.

To my love, Jessica, who came into my life in a period of great darkness giving me strength and support to continue this journey.

I also thank the Federal University of Paraná, the National Council for Scientific and Technological Development (CNPq), and the Higher Education Personnel Improvement Coordination (CAPES) for the financial and infrastructure support that were extremely important and indispensable for the realization of this work.

In addition, I would like to thank my advisors who have become great friends along the way: David Menotti, Yandre Maldonado e Gomes da Costa, for all their attention and help during the development of this research. Thank you not only for the example of extreme professionalism but also for all the extra-professional support.

Here is also my thanks to all the professors who helped me during the development of this research and to the qualification and defense examining committee members for all the notes that were essential for the improvement of this thesis.

Thanks to all the friends and colleagues in the Vision, Robotics and Imaging (VRI) lab with whom I have spent most of my time over the past 5 years. For all the moments of collaboration, exchange of experiences and moments of relaxation during coffees.

“Never look down on anybody unless you’re helping him up”. (Jesse Jackson)

RESUMO

Devido à demanda mundial por sistemas de segurança, a biometria é um tópico crítico de pesquisa em visão computacional. Várias características do corpo humano (impressão digital, face, íris, retina e voz) podem ser usadas como entrada nesses sistemas. Os componentes da região ocular apresentam alto grau de distinção (íris, esclera, toda a área ocular), e a utilização destes em sistemas biométricos proporciona resultados que permanecem entre os mais significativos. Dada a importância dos componentes oculares na biometria, é necessário desenvolver abordagens robustas para a extração da Região de Interesse (RoI) destes, visto que uma área identificada erroneamente pode afetar a eficácia de um sistema inteiro. No entanto, até o momento, nenhum trabalho na literatura aborda as relações contextuais entre os componentes da região ocular. Desta forma, a hipótese deste trabalho é identificar os componentes da região ocular utilizando as informações contextuais da área em que se encontram. Para validar a hipótese proposta toods os experimentos realizados neste trabalho tiveram por finalidade validar: 1) a viabilidade de empregar CNNs para detectar os componentes da região ocular; 2) a complementaridade da detecção simultânea de múltiplos objetos; 3) a viabilidade de realizar segmentação de esclera e íris usando CNNs; 4) o uso de informações contextuais presentes nas imagens como forma de melhorar os resultados apresentados pelas abordagens convencionais de segmentação. Tendo como propósito validar os objetivos definidos, propomos uma abordagem de detecção baseada em YOLO para verificar se a detecção dos componentes da região ocular é possível. Depois disso, avaliamos a complementaridade da detecção simultânea de vários objetos comparando dois métodos, um baseado em *You Only Look Once (YOLO)* e outro baseado *Faster R-CNN + Feature Pyramid Network*. Após o cenário de detecção, todos os nossos esforços foram direcionados à segmentação dos componentes da região ocular. Para avaliar a viabilidade de realizar a segmentação desses componentes, propusemos duas novas abordagens, a primeira baseada em *Generative Adversarial Networks* (GAN) e a segunda baseada em *Fully Convolutional Networks* (FCN). Por fim, para validar o último objetivo apresentado, avaliamos a possibiliade de punir a *Loss* de uma rede neural considerando as informações contextuais presente no conjunto de dados avaliado. Nesse sentido, propomos uma nova abordagem de segmentação baseada em contexto, intitulada *Ocular Region Context Network* (ORCNet), introduzindo uma *loss function* específica, conhecida como *Punish Context Loss* (PC-Loss). A PC-Loss pune o aprendizado de uma rede usando uma diferença percentual entre o *ground truth* e as máscaras de segmentação geradas. A diferença percentual é obtida empregando os conceitos de relacionamento contextual propostos por Biederman, nos quais são avaliados os relacionamentos semântico, espacial e de escala dos objetos presentes em uma imagem. Empregando a Ocular Region Context Network (ORCNet) no conjunto de dados MICHE-I obtivemos resultados promissores nos cenários avaliados (segmentações de íris, esclera e ALL (íris + esclera)), superando os *baselines* utilizados. A ORCNet com ResNet-152 superou o melhor *baseline* (EncNet com ResNet-152) por em 2,27%, 28,26% e 6,43% em termos de $F - Score$, *Error Rate e Intersection Over Union*, respectivamente.

Palavras-chave: Baseado em Contexto, Segmentação, Biometria, Íris, Esclera, Relacionamentos Semântico, Inteligência Artificial Explicável

ABSTRACT

Biometrics is a critical research topic in computer vision due to the global demand for security systems. Various features of the human body (e.g., fingerprint, face, iris, retina, and voice) can be used as input to these systems. Ocular Region Componentss (ORCs) present a high degree of distinction (e.g., iris, sclera, and an entire ocular area), and their use in biometric systems yields the most significant results. Given the importance of ORCs in biometrics, robust approaches for extracting the Region of Interest (RoI) from them are required, as an incorrectly identified area can affect the effectiveness of an entire system. However, to date, no work in the literature addresses the contextual relationships mentioned scenario. To validate the proposed hypothesis, the aims of the experiments in this thesis are to investigate the following: 1) the feasibility of using Convolutional Neural Networks (CNNs) to detect the components of an ocular region, 2) complementarity of the simultaneous detection of multiple objects, 3) feasibility of performing sclera and iris segmentation using CNN-based methods, and 4) use of contextual information present in the images as a technique to improve the results presented by conventional segmentation approaches. To validate the defined objectives, we propose a detection approach based on You Only Look Once (YOLO) to verify the opportunity of ORC detection. Thereafter, we evaluated the complementarity of simultaneous detection of multiple objects by comparing two methods: one is YOLO-based and the other is based on Faster R-CNN + Feature Pyramid Network (FPN). All our efforts were directed toward segmenting the components of the ocular region after the detection scenario investigation. To assess the feasibility of segmenting these components, we proposed two new approaches: based on Generative Adversarial Network (GAN), and based on Fully Convolutional Network (FCN). Finally, to validate the last objective presented, we evaluated the possibility of punishing the loss of a neural network considering the contextual information present in an evaluated dataset. Thus, we propose a new context-based segmentation approach, called Ocular Region Context Network (ORCNet), introducing a specific loss function, known as Punish Context Loss (PC-Loss). PC-Loss punishes the learning of a network by using a percentage difference between the ground truth and generated segmentation masks. The percentage difference is obtained using the contextual relationship concepts proposed by Biederman, wherein the semantic, spatial, and scale relationships of the objects present in a dataset are evaluated. We employed the ORCNet in the Mobile Iris Challenge Evaluation I (MICHE-I) dataset and obtained promising results in the evaluated scenarios (e.g., iris, sclera, and ALL (iris + sclera) segments), surpassing the baselines used. ORCNet outperformed the best baseline (EncNet) by 2.27%, 28.26%, and 6.43% in terms of $F - Score$, Error Rate, and Intersection Over Union, respectively. Keywords: Context-Based, Segmentation, Biometrics,

Iris, Sclera, Semantic Relationships, Explainable Artificial Intelligence.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ASEF	Average of Synthetic Exact Filters
ASM	Active Shape Model
BFPN	Balanced Feature Pyramid Network
BioSec	BioSec
CASIA	Central Asia Student International Academic
CASIA-Interval	Central Asia Student International Academic - Iris -Interval
CASIA-IrisV1	Central Asia Student International Academic - Iris V1
CasiaI3	Central Asia Student International Academic - Iris - Interval v3
CasiaT4	Central Asia Student International Academic - Iris - Thousand v4
CBD-E	Contextual Bidirectional Enhancement
CFPN	Cross-Channel Feature Pyramid Network
CGAN	Conditional Generative Adversarial Network
CLN	Context Learning Network
CNN	Convolutional Neural Network
CrEye-Iris	Cross-Spectral Iris/Periocular
CRF	Conditional Random Field
ED	Encoder-Decoder
EncNet	Context Encoding Network
ER	Error Rate
F1	F-Measure
Fast R-CNN	Fast Region-based Convolutional Network
Faster R-CNN	Faster Region-based Convolutional Network
FCN	Fully Convolutional Network
FDH	Foreground Attentive Detection Heads
FN	False Negative
FP	False Positive
FPN	Feature Pyramid Network
FRGC	Face Recognition Grand Challenge
GAC	Geodesic Active Contours
GAN	Generative Adversarial Network
GPU	Graphical Processor Unit
HCNN	Hierarchical Convolutional Neural Network
HOG	Histogram of Oriented Gradients
IIIT-D CLI	IIIT-Delhi Contact Lens Iris
IITD-1	IITD Iris Image Database 1.0

IJCB	International Joint Conference on Biometrics
IoU	Intersection over Union
IRISSEG	Iris Segmentation Framework
LE-ASM	Local Eyebrow Active Shape Model
Leaky ReLU	Leaky Rectified Linear Unit
mAP	mean Average Precision
MASD	Multi-Angle Sclera Database
MASK-RCNN	Mask Region-Based Convolutional Neural Network
MBGC	Multiple Biometrics Grand Challenge
MER	Mean Error Rate
MFCN	Multi-scale Fully Convolutional Network
MICHE-GS4	MICHE Galaxy S4 Subset
MICHE-GT2	MICHE Galaxy Tab 2 Subset
MICHE-I	Mobile Iris Challenge Evaluation I
MICHE-IP5	MICHE Iphone 5 Subset
MLP	Multi-Layer Perceptron
MobBIO	MobBIO Subset
MobBIOfake	MobBIOfake
MRF	Markov Random Field
MRS	Maximum Radial Suppression
MSRC	Microsoft Research Cambridge
MTM	Markovian Texture Model
NDCCL	Notre Dame Cosmetic Contact Lenses
NDCLD15	Notre Dame Contact Lens Detection 2015
NICE-I	Noisy Iris Challenge Evaluation, Part I
NIR	Near-Infrared
OCAC	Optical Correlation based Active Contours
ORC	Ocular Region Components
ORCNet	Ocular Region Context Network
ORD	Ocular Region Detection
OSIRISv4.1	Open Source Iris Recognition System Version 4.1
PC-Loss	Punish Context Loss
RCNN	Region Based COnvolutional Neural Network
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristics
RoI	Region of Interest
RPN	Region Proposal Network
SBVPI	Sclera Blood Vessels, Periocular and Iris
SSBC	Sclera Segmentation Benchmarking Competition

SVM	Support Vector Machines
TIH	Task-interactive Head
VFC	Vector Field Convolution
VIS	Visible
YOLO	You Only Look Once
YOLO V2	You Only Look Once Version 2

CONTENTS

1 INTRODUCTION

The use of biometrics to verify or identify a person has recently gained popularity because biometric features (physical and behavioral) of a person cannot be lost or forgotten, unlike other forms of identification, such as passwords or identity cards (Bolle et al., 2004; Bowyer et al., 2008).

Several physical characteristics of the human body, such as fingerprints, face, voice, and Ocular Region Componentss (ORCs), can be used as input for biometric systems. ORCs have high discriminative power among the previously mentioned physical attributes, and hence they are an excellent choice for developing a non-invasive user identification system (Das et al., 2013; Zhou et al., 2012; Jain et al., 2016; Wildes, 1997; Zanlorensi et al., 2018; Liu et al., 2016b; Zanlorensi et al., 2019).

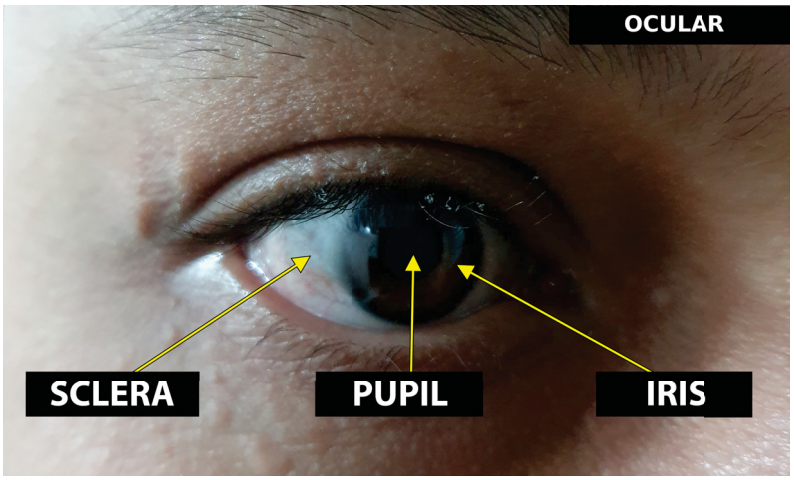
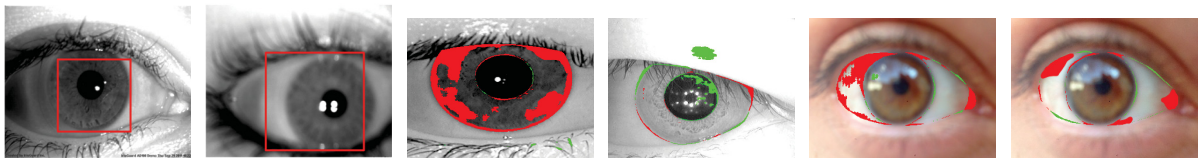


Figure 1.1: Ocular region parts employed in biometric systems.

Figure 1.1 shows the ocular region composed of the iris, pupil, and sclera. According to Bowyer et al. (2008), the pupil region is the central portion of the eye and is generally darker than the iris. However, in some cases, specular reflexes and cataracts may be present, possibly making it lighter. The iris is a colored ring comprising tissues surrounding the pupil through which light enters the eye. The sclera is a white region of connective tissue and blood vessels that surrounds the iris. Meanwhile, there is no standard definition in the literature regarding the location of the Region of Interest (RoI) for the ocular region. Some researchers considered the center of the iris as a reference point and calculated the width and height of theRoI as 6× and 4× the radius of the iris, respectively (Mahalingam et al., 2014; Park et al., 2011; Tan and Kumar, 2013). In contrast, Padole and Proença (2012) proposing using the eye corners as the reference point to calculate the RoI as they are less affected by gaze, pose variation, and occlusion.

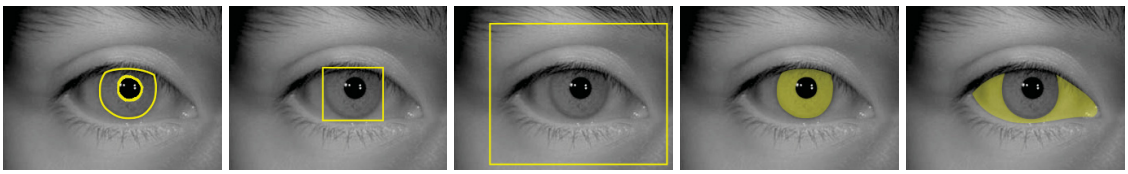
Among the previously mentioned ORCs, the iris presents one of the most accurate results in biometrics since it has a sufficiently complex texture pattern that can be used on the identification task (Liu et al., 2016b; Al-Waisy et al., 2018; Nguyen et al., 2018; Proença and Neves, 2017). However, in recent years, experiments performed using the sclera and entire ocular region have also presented promising results on biometric tasks (Luz et al., 2018; Proença and Neves, 2018; Das et al., 2016, 2017; Delna et al., 2016).



(a) (b) (c)

Figure 1.2: Samples of bad RoIs extraction. In the Figures 1.2(b) and 1.2(c) the green region in images shows wrong segmented regions, while the red region shows not segmented regions.

As said mentioned previously, the ORCs can be employed as input to biometrics once they present a high level of differentiation between users (Wildes, 1997; Jain et al., 2004; Das et al., 2016, 2017; Delna et al., 2016). However, the images must be submitted to a preprocessing stage to extract the RoI. The preprocessing stage is crucial in a biometric system because, if region extraction is erroneously performed, the system effectiveness may be affected (patterns can be removed and/or introduced into the RoI) (Lucio et al., 2018; Rattani and Derakhshani, 2017). Figure 1.2(a) shows wrong detection samples where the RoI is not completely detected and other regions are captured. Similarly, Figures 1.2(b) and 1.2(c) show poor iris and sclera segmentation, with the green and red pixels on the image representing the False Positive (FP) and false negative False Negative (FN), respectively.



(a) (b) (c) (d) (e)

Figure 1.3: Detection and segmentation samples: Figure 1.3(a) present one of the first attempts of iris segmentation, where the RoI were delimited by the internal and external contours of the iris (Liu et al., 2005; Daugman, 2007). Figures 1.3(b) and 1.3(c) present the delimitation approaches employed to detect the iris and eye respectively (Severo et al., 2018; Lucio et al., 2019). Figures 1.3(d) and 1.3(e) present the delimitation approaches employed to segment the iris and sclera respectively (Bezerra et al., 2018; Lucio et al., 2018).

Considering the importance of the preprocessing stage in the ocular region-based biometrics, many different approaches were proposed to solve the RoI extraction problem. The most common delimitation approaches employed in this scenario can be observed in the ground truth samples presented in Figure 1.3.

The most commonly employed approaches as preprocessing steps are contour detection (Liu et al., 2005), Hough transform (Proença and Alexandre, 2006), active contours (Ouabida et al., 2017; Shah and Ross, 2009), integro-differential equation (Tan et al., 2010), Maximum Radial Suppression (MRS) (Podder et al., 2015), Markovian Texture Models (MTMs) (Haindl and Krupička, 2015), Convolutional Neural Networks (CNNs) (Lucio et al., 2018; Bezerra et al., 2018; Jalilian et al., 2017; Liu et al., 2016a; Lucio et al., 2019; Zanlorensi et al., 2020).

Despite many proposed preprocessing approaches, none of them work correctly in non-controlled environments because the input image contains a portion of the face and possibly other types of objects (e.g., background objects, glasses, nose, and mouth). In addition to the previously described issue, none of the preprocessing approaches address RoI extraction in the context where they are used.

Thus, the main question addressed in this study is “Can we improve the RoI extractor in ORCs considering the context information present in an image?”

To answer this question, we propose the Ocular Region Context Network (ORCNet) as a new segmentation approach for the simultaneous segmentation of the sclera and iris using contextual information. The proposed approach employs the segmentation and detection knowledge obtained in the first stages of this work combined with the contextual relationship concept proposed by Biederman (1972). The following were the evaluated contextual relationships:

- **semantic:** evaluates the likelihood of an object being found in some scenes but not in others;
- **spatial:** evaluates the placement among objects in a scene;
- **scale:** evaluates the size ratio among the different classes of objects in a scene.

1.1 MOTIVATION

Biometrics refers to the use of physiological and behavioral characteristics of humans for personal identification (Das et al., 2013). Such characteristics are particularly important since they cannot be changed, forgotten, lost or stolen, providing an unquestionable relationship between the individual and the application that makes use of them (Menotti et al., 2015).

From the physiological attributes, the iris, sclera, the entire ocular region deserve special attention since they have a sufficiently complex texture pattern that can be used for the identification task, becoming an excellent option for developing a non-invasive biometric system (Wildes, 1997; Zanlorensi et al., 2018; Liu et al., 2016b; Al-Waisy et al., 2018; Nguyen et al., 2018; Proença and Neves, 2017).

Typically, in ORC-based biometrics, RoI extraction is the first step wherein efforts should be applied to a reliable recognition system. Because incorrect extraction of a region can reduce or introduce new patterns (e.g., eyelashes, eyelids and unnecessary portion of skin) in the input data, compromising the effectiveness of a biometric system (Lucio et al., 2018, 2019; Bezerra et al., 2018; Severo et al., 2018).

Many studies were conducted using CNN approaches to solve the previously presented problem (Lucio et al., 2018, 2019; Bezerra et al., 2018; Jalilian et al., 2017; Liu et al., 2016a; Severo et al., 2018). The primarily reason is that deep learning has shown promising results in many other correlated areas (e.g., face recognition, natural language processing, and speech recognition). Despite that numerous studies have been conducted on the preprocessing stage of biometric systems, no studies have handled the contextual information in images.

Finally, the importance of iris-based biometrics and other ORC-based emerging approaches such as periocular and sclera-based are considered. The aim to extract the RoI as perfectly as possible becomes the reason for developing new extraction approaches. In this manner, we notice the feasibility of exploring this problem. Thus, the main problem that we considered in this study is “how to create a robust RoI extractor for the ORC that considers the context where each of these lies, regardless of the scenario.”

1.2 PROBLEM DEFINITION

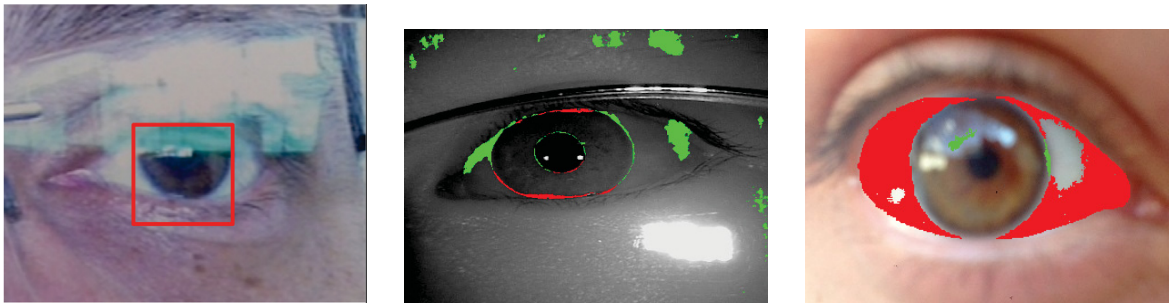
Ocular region component-based biometrics is a branch of biometric recognition technology that exploits various eye characteristics for identity inference. Biometrics based on the iris and retina, which are the ocular regions, have produced the most accurate scores (Das et al., 2013). However, these biometric systems require, user collaboration and an intrusive image acquisition scheme, respectivelly (Das et al., 2015). In addition to the iris and retina biometric features, the eye has a

white region around the eyeball, known as the sclera, that contains a pattern of blood vessels that can be used for personal identification (Das et al., 2016, 2017; Delna et al., 2016).

Commonly, in ORC-based biometrics, a preprocessing step is necessary since the high reliability presented by these approaches is conditioned on their input data. In this manner, the extraction of the Region of Interest (RoI) is the first step in which efforts should be applied to an ORC-based recognition system.

(a) Full face with specular highlights on iris (b) Eyeglasses with specular highlights (c) Eyeglasses and background in image

Figure 1.4: Samples of difficult images



(a) (b) (c)

Figure 1.5: Bad iris and sclera RoI extraction samples. In (a) and (b), the image's high amount of specular highlights has affected the iris detection and segmentation, respectively. In (c) beyond the specular highlights the low resolution of the input image has affected the sclera segmentation.

Tipically, the input image of ocular-based biometrics also has parts of the face and even other types of objects (e.g., background objects, glasses, nose, and mouth; Figure 1.4). This makes the RoI extraction task even more challenging since the currently available approaches do not work correctly, assigning unwanted regions as belonging to the RoI Figure 1.5. Finally, the main problem that we considered is the ORC extraction because the previously presented problems (see Figure 1.4) can not affect the extracted RoI.

1.3 CHALLENGES

Considering the mentioned problems, the main challenge evaluated in this study is how to extract the RoIs from the ocular region using the contextual information on the area by employing deep learning techniques. Therefore, we present the specific objectives as follows:

- To annotate the segmentation datasets used in a biometrics scenario based on the ocular region, since there are no specific datasets available so far for the segmentation/detection task of the components of the ocular regions are not available thus far;
- To design an approach that can understand the contextual information of the images from different ocular biometrics sources, such as near-infrared and visible images from constrained and unconstrained environments.

1.4 HYPOTHESES

However, although several published studies are addressing the ORC delimitation, they did not consider contextual information present in the images. Therefore, the hypothesis of this work is that **it is possible to extract ORCs can be extracted by considering the contextual information of the area where this information it is found**? We also investigate the hypothesis that, in this scenario, the presented results can equate or exceed those obtained using conventional methods of detection and segmentation. Considering the unexplored gap in RoI extraction and the importance of this step in biometrics, as mentioned previously, we investigate the following topics regarding iris and sclera detection/segmentation using deep learning techniques:

- Impact of ocular region detection as a first step on sclera segmentation;
- Effects of the use of coarse annotations on iris and sclera detection;
- Complementarity of simultaneous detection of the iris and sclera;
- Benchmark CNN architectures on ORC detection/segmentation;
- Understanding how the image elements were related contextually.

1.5 OBJECTIVES

Considering the hypothesis, we define the main objective:“to propose a robust and context-based ORC segmentation approach by utilizing CNNs”. The aim is to identify the best architecture that can be used alone or with already known types of architecture to solve the problem. Further, the new approach should be versatile and effective. “Versatile” means that a context-based detection method that can be used in the broadest variety of domains possible. “Effective” means an architecture that presents results that are close to or better than those obtained using methods conventionally employed in this type of task. The general objective can be achieved by accomplishing the following specific aims:

1. Identify the most relevant and available CNN-based detection approaches available in the literature, and evaluate their performance in Ocular Region Components (ORC) detection;
2. Investigate whether a multi-class ORC object detection approach can outperform a single detection method;
3. Identify the most relevant available CNN-based segmentation approaches in the literature, and assess their performance in detecting ORC;

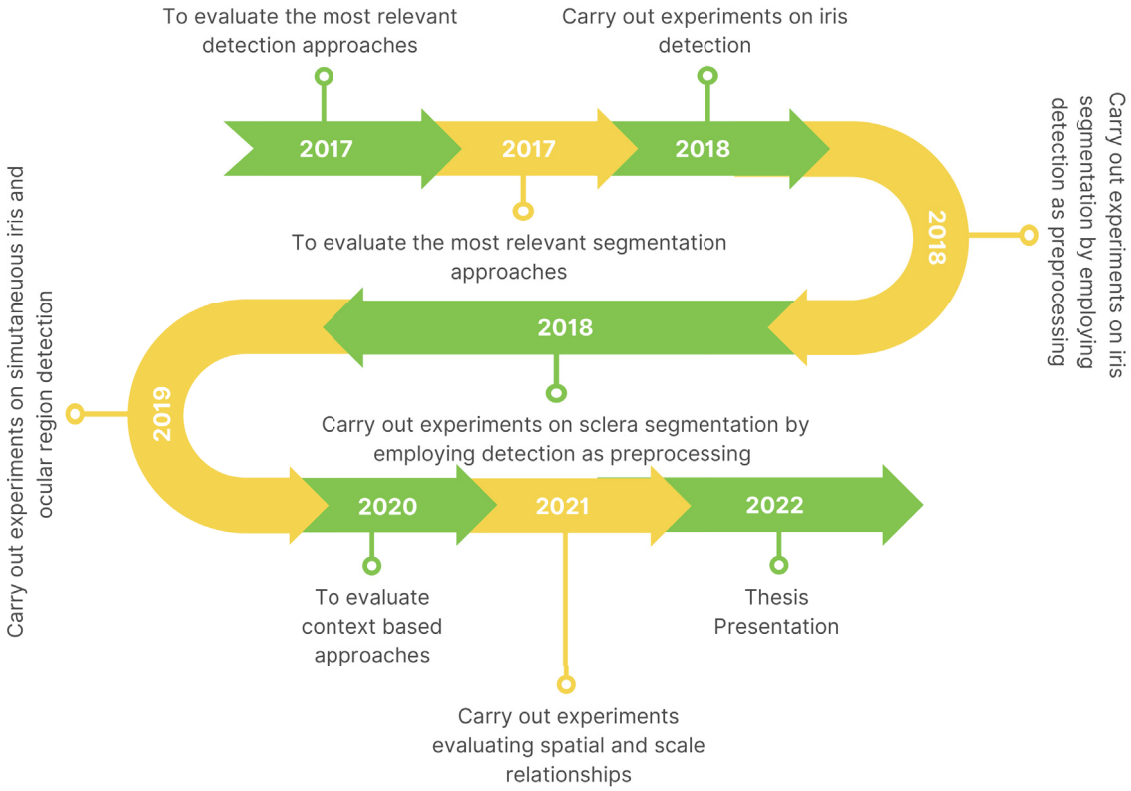


Figure 1.6: Workflow used to address the objectives presented in this section.

4. Present a new CNN model that can learn to distinguish all the ORC based on the context in which they are.

To achieve all the these objectives, a workflow was developed (Figure 1.6). We can observe the performed activities in chronological order since the beginning of this study.

1.6 CONTRIBUTIONS

To achieve these objectives, some methodologies, and approaches are investigated. Other approaches are obtained:

- Review and investigation of the state-of-the-art methods using a CNN, to segment and detect the ORC;
- Two published papers regarding segmentation using Fully Convolutional Network (FCN) and Generative Adversarial Network (GAN) to segment ORCs. The focus of the first study is on the sclera region (Lucio et al., 2018), but the second one is on the iris, which has attracted our attention (Bezerra et al., 2018);
- Evaluation and comparison of different approaches for detection and segmentation that were previously employed in other domains (e.g.,scene understanding and face detection);

- Annotation of more than 170, 000 images from the well-known iris datasets, which were employed in the detection tasks of this study;
- Annotation of more than 10, 000 images from well-known iris datasets, which were employed in the segmentation tasks of this study;
- A context-based approach for ORCs RoI extraction that achieved state-of-the-art results.

1.7 LIST OF PUBLICATIONS

The following are some of the international prizes were obtained by the mentioned studies:

1.7.1 Published Own Productions

- **Simultaneous iris and ocular region detection using coarse annotations**; View article; *32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*; 2019 (Lucio et al., 2019).
- **Fully convolutional networks and generative adversarial networks applied to sclera segmentation**; Diego R. Lucio, Rayson Laroca, Evair Severo, Alceu S Britto, David Menotti; *9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*; 2018 (Lucio et al., 2018).

1.7.2 Collaborative Productions

- **On the Cross-dataset Generalization for License Plate Recognition**; Rayson Laroca, Everton V Cardoso, Diego R Lucio, Valter Estevam, David Menotti; *arXiv preprint*; 2022 (Laroca et al., 2022).
- **SSBC 2020: Sclera segmentation benchmarking competition in the mobile environment**; M Vitek, A Das, Y Pourcenoux, A Missler, C Paumier, S Das, I De Ghosh, Diego Rafael Lucio, Luiz A Zanlorensi, David Menotti, Fadi Boutros, Naser Damer, Jonas Henry Grebe, Arjan Kuijper, J Hu, Y He, C Wang, H Liu, Y Wang, Z Sun, D Osorio-Roig, Christian Rathgeb, Christoph Busch, J Tapia, Andres Valenzuela, G Zampoukis, L Tsochatzidis, Ioannis Pratikakis, S Nathan, R Suganya, Vineet Mehta, Abhinav Dhall, K Raja, G Gupta, JN Khiarak, M Akbari-Shahper, F Jaryani, Meysam Asgari-Chenaghlu, R Vyas, S Dakshit, P Peer, Umapada Pal, V Štruc; *2020 IEEE International Joint Conference on Biometrics (IJCB)*; 2020 (Vitek et al., 2020).
- **UFPR-Periocular: A Periocular Dataset Collected by Mobile Devices in Unconstrained Scenarios**; L. A. Zanlorensi, R. Laroca, D. R. Lucio, L. R. Santos, A. S. Britto Jr., D. Menotti; 2020 (Zanlorensi et al., 2020).
- **CNN hyperparameter tuning applied to iris liveness detection**; Gabriela Y Kimura, Diego R Lucio, Alceu S Britto Jr, David Menotti; 2020 (Kimura. et al., 2020).
- **Efficient Deep Learning Model for COVID-19 Detection in large CT images datasets: A cross-dataset analysis**; Pedro Silva, Eduardo Luz, Guilherme Silva, Gladston Moreira, Rodrigo Silva, Diego Lucio, David Menotti; *Informatics in Medicine Unlocked*; 2020 (Silva et al., 2020).

- **Deep representations for cross-spectral ocular biometrics**; Luiz A Zanlorensi, Diego Rafael Lucio, Alceu de Souza Britto Junior, Hugo Proença, David Menotti; *IET Biometrics*; 2020 (Zanlorensi et al., 2020).

- **Robust iris segmentation based on fully convolutional networks and generative adversarial networks**; Cides S Bezerra, Rayson Laroca, Diego R Lucio, Evair Severo, Lucas F Oliveira, Alceu S Britto, David Menotti; *31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*; 2018 (Bezerra et al., 2018).

1.7.3 Under Review

- **Federated Learning Enables Big Data for Rare Cancer Boundary Detection**; Pati et al.; (*Nature Medicine*). Submitted on: 23/03/2022.

- **Exploring Bias in Sclera Segmentation Models: A Group Evaluation Approach**; Vitek et al.; (*IEEE Transactions on Information Forensics and Security*). Submitted on: 21/03/2022.

- **ORCNet: A context-based network to simultaneously segment the ocular region components**; Diego Rafael Lucio, Luiz A Zanlorensi, Yandre Maldonado e David Menotti. (*Neural Processing Letters*). Submitted on: 01/04/2022.

- **Pupil Constrictions and Dilations Effects as Data Augmentation on an Iris Recognition CNN Approach**; Diego Rafael Lucio, Luiz A Zanlorensi, Yandre Maldonado e David Menotti. (*The 33rd IEEE International Conference on Tools with Artificial Intelligence*). Submitted on: 05/04/2022.

1.7.4 International Prizes

Taking into account the research presented in this thesis some international prizes were conquered:

- **1st Place** on 2020 IEEE WCCI VISOB 2.0 Competition.

  Info: The VISOB 2.0 competition aimed to evaluate and compare the performance of ocular biometrics recognition approaches in visible light using (a) stacks of five images captured in burst a mode and (b) subject-independent evaluation, where subjects do not overlap between training and testing set.

- **2nd Place** on SSBC 2020: Sclera Segmentation Benchmarking Competition in the Mobile Environment

  Info: Sclera biometrics have gained significant popularity among emerging ocular traits in the past few years. To establish the sclera as a viable biometric trait and to evaluate its potential for automated biometric systems, several works have been presented in the literature that employed sclera both individually and with the iris. Despite these initiatives, sclera biometrics should be investigated more extensively to ascertain their usefulness. Among the numerous challenges that still exist in this filed, efficient and robust sclera segmentation is one of the most important ones, affecting all downstream tasks from normalization to recognition. While a considerable numberf of research has been directed towards sclera segmentation in recent years, the performance of existing techniques especially in cross-sensor scenarios and across different acquisition conditions is still not well explored. To investigate these issues, to document recent

developments and to attract (and raise) the interest of researchers in this emerging biometric trait, the Sclera Segmentation Benchmarking Competition (SSBC) 2020 was proposed - a competition (and group benchmarking effort) held in conjunction with International Joint Conference on Biometrics (IJCB) 2020 focusing on the problem of sclera segmentation.

1.8 DOCUMENT ORGANIZATION

This work is further organized into 6 chapters. Chapter 2 presents the theoretical foundation for detection and segmentation approaches. The first section presents the most important methods of the detection scenarios, while in the next one provides the most known methods employed in the segmentation scenario are presented. Chapter 3 describes the theoretical foundation employed in this study. The first section presents the essential methods for the detection scenarios, the second section discusses the most well-known methods employed in the segmentation scenario, and the final section presents the concepts of RoI extraction based on contextual relationships. Chapter 4 describes the well-known databases of iris and periocular region available in the literature. The proposed methodologies and the obtained results are detailed in Chapter 5. Finally, the conclusion and future studies are presented in Chapter 6.

2 THEORETICAL FOUNDATION

In this chapter, the most common CNN based approaches used to detect and segment objects in the images are presented. Section 2.1 and 2.2 describes the CNN concepts respectively. Then Section 2.3 shows the object detection approaches. Finally, Section 2.4 describes the segmentation approaches.

2.1 DEEP LEARNING

The performance of a classification system strongly depends on the choice of set of characteristics (representation) used (Bengio et al., 2013). Thus, much effort on the development of machine learning applications is aimed to improve the pre-processing and data transformation steps, to obtain a set of features with high discrimination power.

In general, this feature representation is empirically defined by a human expert who, with prior knowledge of the problem, judges which features have greater power of discrimination for a given application. Consequently, the extraction of such characteristics is carried out using a certain method.

In this context, the automatic learning of the representation (without the need of a human expert) can facilitate the extraction of important information in the construction of classification, detection, or recognition systems. Among the forms of learning, we can mention representation learning, which involves a set of methods wherein the representations necessary for the detection or classification are automatically discovered from raw data (LeCun et al., 2015).

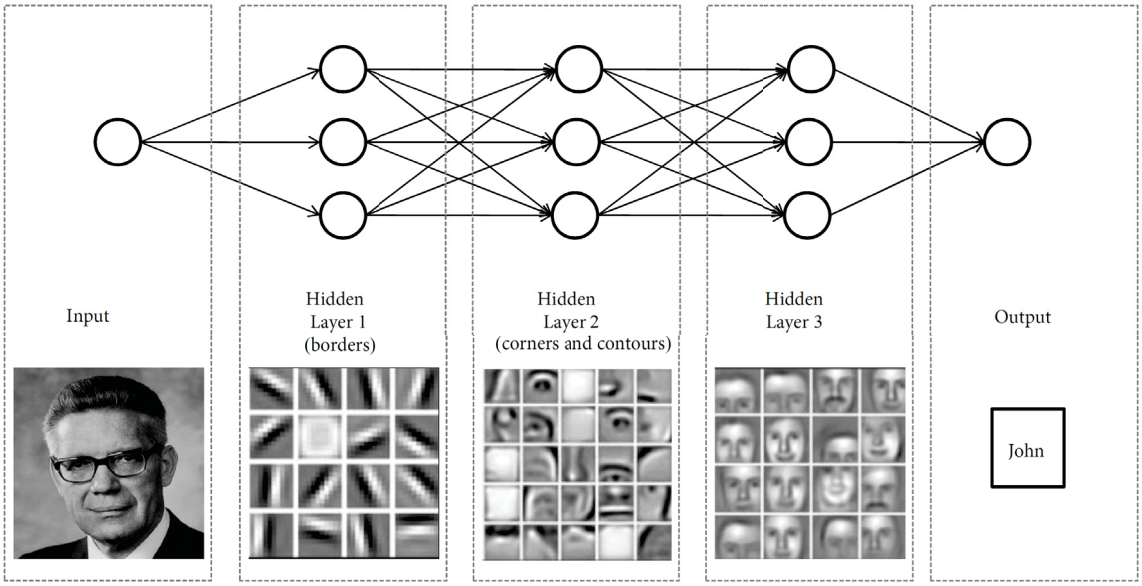


Figure 2.1: Illustration of a deep learning model.

The central problem in Representation Learning is that it can be very difficult to extract high-level abstract features from raw data. However, deep learning solves this problem by introducing representations that are expressed in terms of other, simpler representations (Goodfellow et al., 2016).

As shown in Figure 2.1, in the first representation layer, features referring to the presence or absence of edges in specific orientations and locations in the image are extracted. Then corners and contours are detected on the second layer. The third layer is where parts of objects are found, locating specific sets of contours and corners. Finally, subsequent layers would detect specific objects as combinations of the parts found in the previous layer (Goodfellow et al., 2016).

According to the observation made by LeCun et al. (2015), one of the main features of deep learning is that the layers of representations are learned through a general-purpose learning procedure and, therefore, requires minimal human interference.

Initially, deep learning approaches were mainly applied to the handwritten digit recognition problem, surpassing the results obtained by a Support Vector Machines (SVM) based approaches. As a result, deep learning was adopted in approaches for recognizing objects in natural images. An example is the progress achieved by (Krizhevsky et al., 2012) in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), reducing the error rate of the state-of-the-art approaches from 26.2% to 15.3% (Bengio et al., 2013).

2.2 CONVOLUTIONAL NEURAL NETWORKS

An example of an approach based on deep learning is the CNN, which is a concept introduced by (Fukushima and Miyake, 1982) and popularized by (LeCun et al., 2012). A CNN is a variation of a Multi-Layer Perceptron (MLP) and comprises layers with different functions. These networks can learn representations based on a set of training samples. CNNs have currently been used to improve state-of-the-art results in several image recognition and/or classification problems. These include, biometric approaches (Lucio et al., 2018; Bezerra et al., 2018), systems security and monitoring, and license plate recognition (Laroca et al., 2018, 2021), among others (LeCun et al., 2015; Ahuja et al., 2017; Dumoulin and Visin, 2016; Krizhevsky et al., 2012).

In general, the examples are initially applied to the input layers, known as convolutional layers (Vargas et al., 2016). These layers are composed of neurons, and each neuron is responsible for applying a filter to a specific area of an image. Basically, a neuron is connected to a set of pixels from a previous layer and each connection is assigned a weight. The respective weights of their connections produce an output that is passed to the next layer. The weights assigned to the connections of a neuron can be interpreted as a matrix representing the filter of a convolution of images in the spatial domain (kernel). These weights are shared between neurons of the same layer, allowing the filters to learn patterns that occur in the image.

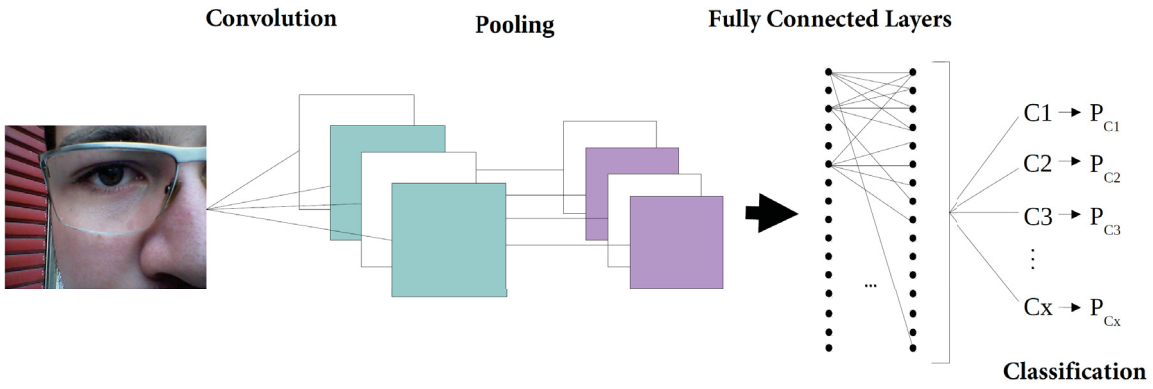


Figure 2.2: Illustration of a deep learning model.

There are typically four basic types of layers in a CNN: convolution layer, rectified linear unit layer, pooling layer (aggregation) and classification layer (Fully Connected Layer). Figure 2.2 illustrates the organization of these layers.

Convolution: The first layer comprises a layer of convolutional type . A filter (also called a kernel) with a set of weights is slid across the image and the result of this convolution is updated in the feature map (Dumoulin and Visin, 2016). Each filter can be considered a descriptor of features, such as edges, lines, curves, colors, and others.

Rectified Linear Unit (ReLU): Basically, the ReLU layer is aims to introduce non-linearity into the network, since most data in the real world are non-linear (and the convolution operation is linear). This layer performs this function through an operator, which basically replaces all negative values in the activation map with zero. In addition, ReLU solves the vanishing gradient problem (vanish gradient).

Pooling: This layer works similarly to convolution, making use of a kernel that applies a certain calculation to the values of each feature map. Its objective is to reduce the dimensionality of each feature map, trying to preserve the most relevant information. Consequently, this technique reduces the number of parameters to be learned in a network, contributing to overfitting control. The kernel-enforced function can be implemented in several ways. The maximum, average, and sum functions are some examples of possible functions to be used in the Pooling layer.

Classification (Fully Connected Layers): This layer comprises a MLP, with a softmax activation function in the output layer, guaranteeing the estimation of as afterthought probabilities. Thus, the output generated through the convolution+ReLU+Pooling layers represent high-level features that serve as input to a MLP. Furthermore, this representation can be used with any other classifier.

2.3 OBJECT DETECTION

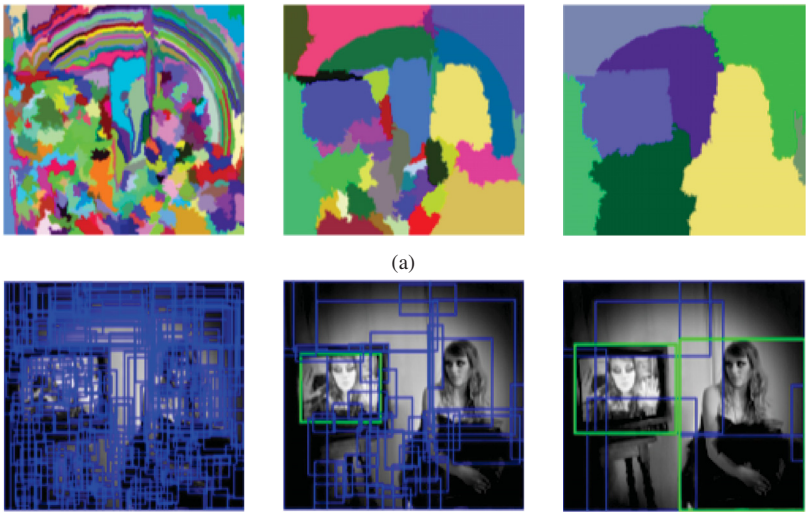
In this section, the CNN-based object detection approaches are investigated. Section 2.3.1 explores the region-based CNN and Section 2.3.2 investigates the You Only Look Once (YOLO) based segmentation approaches.

2.3.1 Region-Based Convolutional Networks

This section highlights the most important studies that use region-based convolutional networks to detect objects in images. Uijlings et al. (2013) were the first to explore this scenario by presenting a *selective search* approach. Subsequently, several studies have presented new techniques to detect objects considering the region-based concept. After Uijlings et al., many other authors have shown new techniques to detect objects considering the region-based concept.

Among the proposed methods, some deserve to be highlighted, namely: Region Based COnvolutional Neural Network (RCNN) (Girshick et al., 2014), Fast Region-based Convolutional Network (Fast R-CNN) (Girshick, 2015), Faster Region-based Convolutional Network (Faster R-CNN) (Ren et al., 2015), and Mask Region-Based Convolutional Neural Network (MASK-RCNN) (Ren et al., 2015).

In an object detection scenario, Uijlings et al. (2013) developed the *Selective Search* an approach that combines exhaustive search and segmentation to perform the RoI extraction. The proposed method detects an object by initializing small regions in the image and merges them by using a hierarchical grouping. The parts are merged according to a variety of color spaces and similarity metrics. Thus the output presented by the method is regions proposals which



(a)

(b)

Figure 2.3: Sample of selective search : 2.3(a) segmentation result of the algorithm; 2.3(b) proposed regions of the algorithm (Uijlings et al., 2013).

could contain an object. Figure 2.3 shows a sample of the segmentation and the proposed regions detected using this approach.

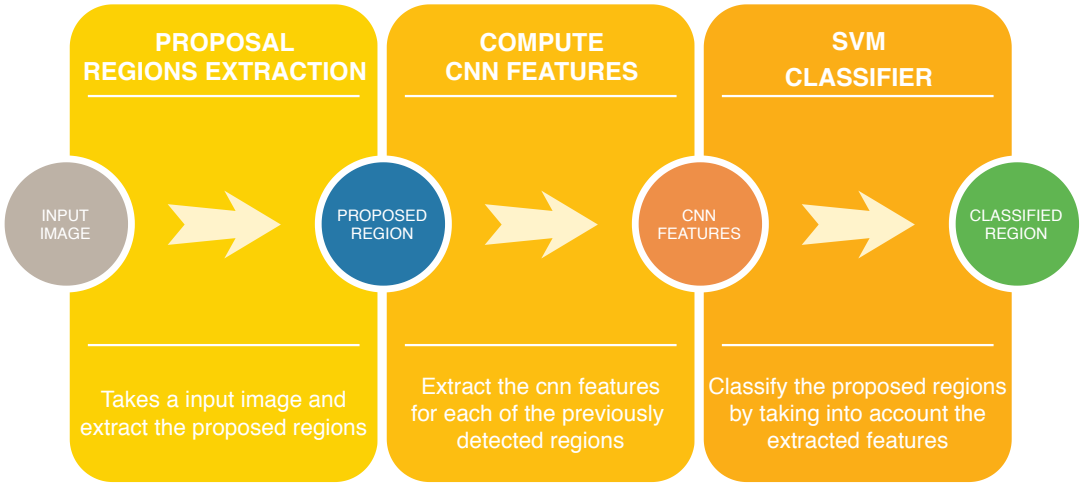


Figure 2.4: Region-based Convolutional Network object detection overview.

Girshick et al. (2014) have developed the RCNN an approach to object detection that combines deep learning with the selective search proposed by Uijlings et al. (2013). The proposed regions identification was made by using the selective search method, and after this, a set of features are extracted for each one of these regions by using a CNN model. By using the extracted features, the object location is made using the Support Vector Machines (SVM) classifier. Figure 2.4 shows an overview of the RCNN steps to detect an object.

Trying to reduce the time consumption of the RCNN, Girshick (2015) presented the Fast R-CNN. Figure 2.5 shows the work-flow employed in the proposed approach, in this a CNN is used to process an entire image, and generate feature maps, that will be used to locate the bounding boxes by using the selective search method. In the next step for each RoI a feature vector is extracted using fully connected layers, and after the class probabilities and the Bounding

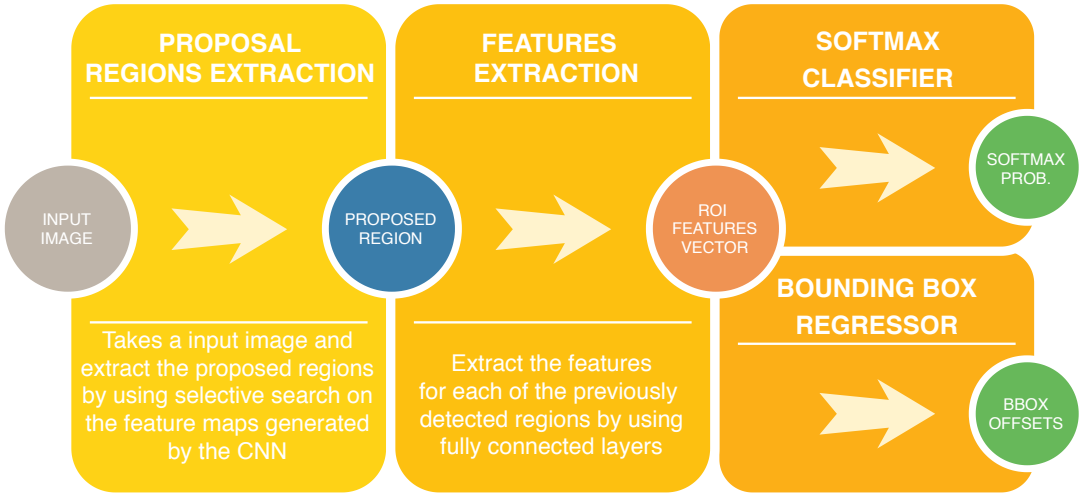


Figure 2.5: Fast Region-based Convolutional Network object detection overview.

Boxes (BBOX) offsets is obtained employing a *Softmax Classifier* and a *Bounding-Box Regressor* respectively.

Taking into account the high computational power required and the time spent to detect the regions of interest using the selective search method, Ren et al. (2015) have proposed the Faster R-CNN a new approach to detect object that combines the Region Proposal Network (RPN) and the Fast R-CNN techniques.

The RPN directly generate region proposals, predict bounding boxes and detect objects, taking as input feature maps and applying over these a 3×3 sliding window, and outputs a feature vector linked to two fully-connected layers (RoI Pool), one for bounding boxes regression and another to bounding box classification.

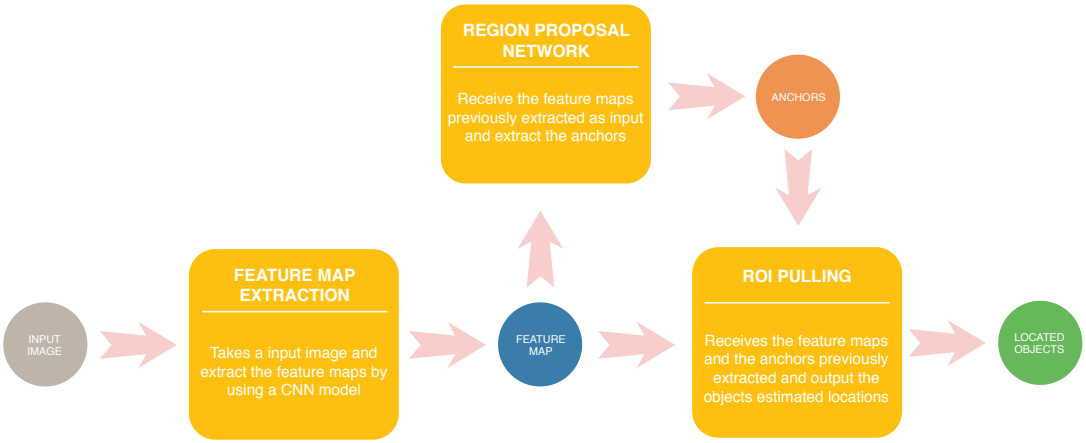


Figure 2.6: Faster Region-based Convolutional Network object detection overview.

The Faster Region-based Convolutional Network (Faster R-CNN) takes as input an image and produces feature maps which are used as input by the RPN which identify the proposed regions. Since many regions are proposed, maximum k regions are defined, and the output of the box-regression layer of the RPN has a size of $4k$ (coordinates, height, and width). By using the k proposed regions, the box-classification layer have a length of $2k$ "objectness" scores which inform if the proposed box has an object or not. The k proposed regions (anchor boxes) identified feeds a Fast R-CNN model which predict the object's location. Figure 2.6 present a Faster R-CNN work-flow described in this paragraph.

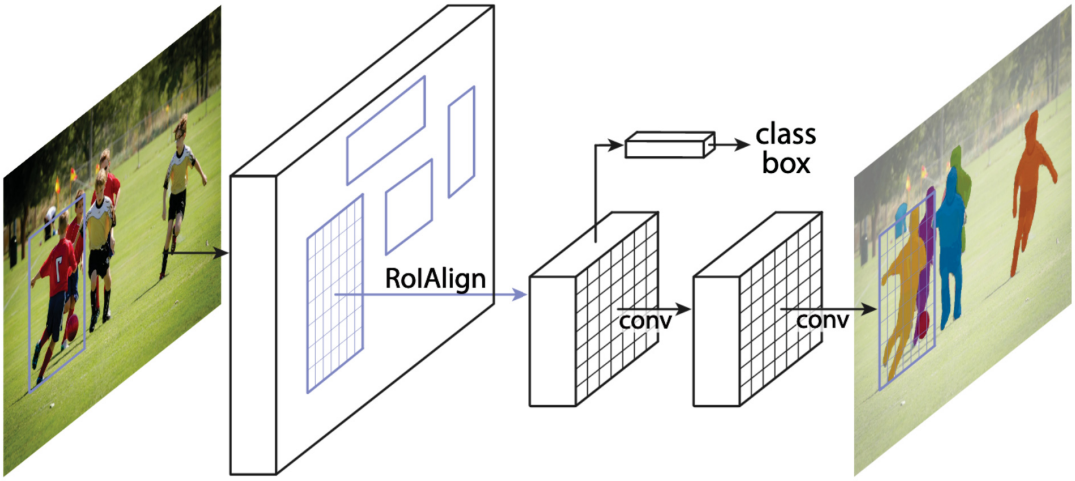


Figure 2.7: MASK-RCNN architecture overview adapted from (He et al., 2017).

He et al. (2017) presented an extension of the Faster R-CNN. In this is added a parallel branch to the bounding box detection to predict object mask.

The MASK-RCNN uses the Faster R-CNN pipeline with three output branches, and for each candidate object the RPN is used to predict: a class label, a bounding box offset and the object mask. Trying to remove the quantization of the coordinates and compute the exact values of locations present in Faster R-CNN, a RoI Align layer replace the RoI Pool. The new layer provides scale and translation equivariance with the region proposals.

Figure 2.7 shows an overview of the MASK-RCNN architecture, in this an image feeds a ResNeXt network with 101 layers which is responsible for detecting the RoIs, and each detected RoI is processed by the RoI Align layer, which generates feature maps. The output of the RoI Align layer, is the input of two branches, the first is a fully-connected layer which computes the coordinates of the bounding boxes and the probabilities associated to the objects, and the second computes the mask of the detected object.

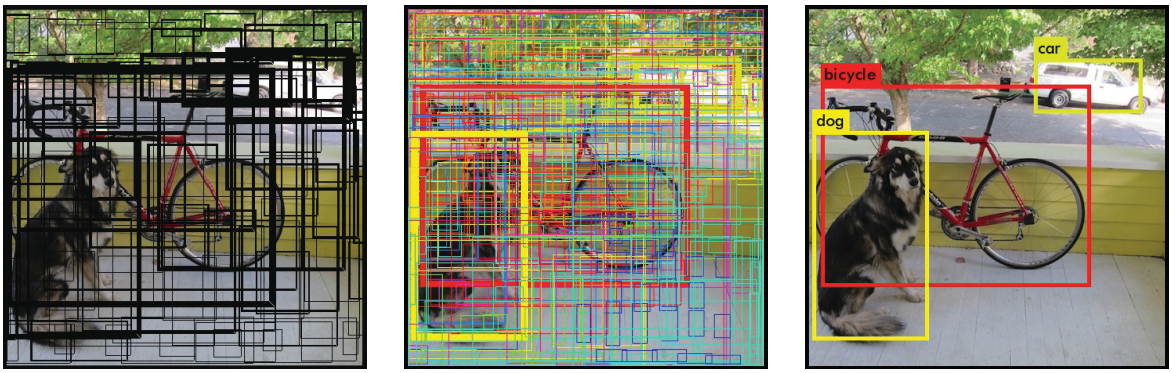
The main particularity of the Mask R-CNN model is the use of a multi-task loss combining the losses of the bounding box coordinates, the predicted class, and the segmentation mask, trying to solve complementary tasks leading to better performances on each job.

2.3.2 You Only Look Once

In pattern recognition, object detection pipelines usually start by extracting features from input images in a sliding window or in some subset of regions in the image. Then the classifiers are used to identify objects in the feature set (Redmon et al., 2016a).

Another possibility is to make use of a deep learning architecture (VGGNet (Simonyan and Zisserman, 2014) or GoogLeNet (Szegedy et al., 2015), for example) originally used for image classification and transform it into an object detector. In this case, a sliding window would be passed through the image. Using a sliding window provides hundreds or thousands of possible predictions for this image and the ones that the classifier reports the highest probability will be considered. This approach works, however it will be very slow as it is necessary to run the classifier numerous times.

A more efficient approach is first to predict which parts of the image contain information of most interest (the so-called Region Proposals) and then run the classifier only on those regions.



(a) (b) (c)

Figure 2.8: YOLO object detection from (Redmon et al., 2016a)

Thus, the classifier will need to perform less work than sliding a window over the entire image numerous times (Girshick, 2015; Ren et al., 2015).

Unlike techniques based on region or sliding window proposal, YOLO (Redmon et al., 2016a) consists of an approach that reformulates object detection as a single regression problem, starting from image pixels to bounding boxes coordinates and class odds. Thus, YOLO globally analyzes the image in the detection process. As its name (You Only Look Once) implies, the network analyzes the image only once, but in an intelligent way (Redmon et al., 2016a).

In general, the image is divided into cells that are responsible for defining bounding boxes (Figure 2.8(a)). Bounding boxes involve possible objects to be detected. By defining confidence values (generated for each bounding box) and class probabilities for each cell, the probability of a given region containing a specific type of object is measured (Redmon et al., 2016a).

For example, the thick yellow bounding box, illustrated in Figure 2.8(b), would present a high probability that there is an object of the dog class (if the network has been trained for this) in its limits. An example of what the final object detection would look like can be seen in Figure 2.8(c).

It is important to note that, to carry out this detection process, the neural network analyzes the image only once, which is why YOLO detects objects with greater speed. Like other CNNs, YOLO is made up of three layers of main operations for object detection, which are: Convolution, Max Poollig and Classification, which occurs through fully connected layers.

As can be seen in Figure 2.9, YOLO has 24 convolutional layers followed by 2 fully connected layers. The first 20 convolutional layers followed by a pooling layer and a fully connected layer were used for pre-training the classification task on the ImageNet database. Then four convolutional layers and two fully connected layers were added for detection. A linear activation function is used for the final layer and all other layers use Leaky Rectified Linear Unit (Leaky ReLU). In addition, the network input resolution has been increased from 224×224 to 448×448, as detection often requires fine-grained information. Finally, there is a 0.5 rate dropout layer after the first connected layer to avoid co-adaptation between layers (Redmon et al., 2016a).

There is also a variation of the original YOLO version called Fast-YOLO (or Tiny-YOLO). Compared to the original version, all training and testing parameters are the same. However, Tiny-YOLO uses fewer convolutional layers (9 instead of 24) and fewer filters on those layers.

Despite Tiny-YOLO having a considerably smaller architecture, it is still able to detect objects very accurately and more quickly. In Figure 2.10, the detections obtained in the same image with Tiny-YOLO and YOLO are shown, using standard parameters.

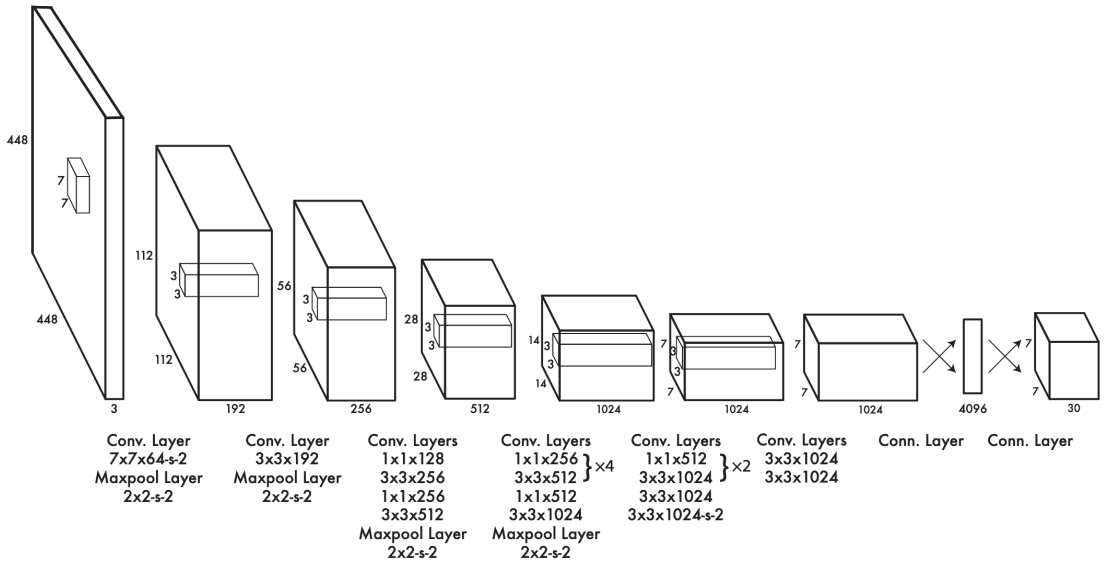
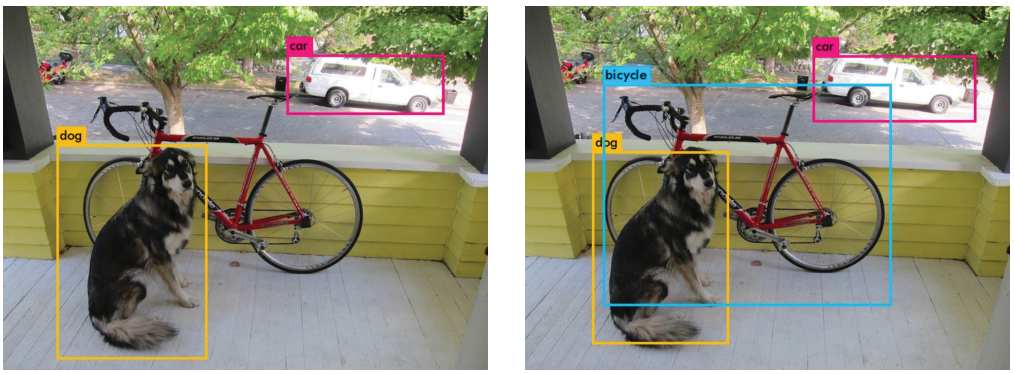


Figure 2.9: YOLO architecture. It has 24 convolutional layers followed by 2 fully connected layers. Alternating 1×1 convolutional layers reduce the feature space of previous layers. Convolutional layers are pre-trained at half resolution (224×224) and then at double resolution for detection. Image reproduced from (Redmon et al., 2016a).



(a) (b)

Figure 2.10: Example of object detection with (a) Tiny-YOLO and (b) YOLO. Replicated image from https://pjreddie.com/darknet/yolo.

In face of the promising results achieved, YOLO has some limitations. The first limitation observed is that each grid cell predicts, by default, only two bounding boxes and can only have one class. This limits detection in regions of the image where objects are very close, as illustrated in Figure 2.11(b).

Another observation to be made is that errors in small and large bounding boxes are treated equally in the loss function. However, a small error in a large bounding box may not be significant, but a small error in a small bounding box has a much greater effect on the Intersection over Union (IoU) metric (Redmon et al., 2016a).

As reported by Redmon and Farhadi (2017a), YOLO has higher location errors and lower recall when compared to some detection systems proposed later. Taking into account the shortcomings of the initial version of YOLO (Redmon and Farhadi, 2017a) proposed the YOLOv2 aiming to significantly improve its accuracy and make it faster.

A new classification model, called Darknet-19, is used as the basis for YOLOv2. Such a model has 19 convolutional layers and 5 max-pooling layers. An improvement of more than 2% in

(a) (b)

Figure 2.11: Illustration of one of YOLO's limitations, where detection can miss objects that are too close. In (a), there are 9 individuals in the lower left corner, but only 5 were detected, as can be seen in (b). Replicated image from (Redmon et al., 2016a).

mean Average Precision (mAP) was achieved by adding batch normalization on all convolutional layers. In this way, the dropout layer was removed without overfitting. In addition, an increase of almost 4% of mAP was achieved through the classification in high resolution, that is, after training the classification network on images of 224×224 pixels, the network was adjusted to the resolution of 448×448 in ImageNet (Redmon and Farhadi, 2017a).

YOLO's fully connected layers have been removed and YOLOv2 uses anchor boxes to define the bounding boxes. Thus, a class and the probability that the region contains some object are predicted for each anchor box (Redmon and Farhadi, 2017a). As an example, consider that 5 anchor boxes with specific proportions are created, as illustrated in Figure 2.12. Instead of predicting 5 bounding boxes, YOLOv2 predicts trade-offs for each of these anchor boxes. In this way, the network just adjusts the size of the anchor closest to the size of the object. This feature, according to the authors, simplifies the problem and facilitates network learning. YOLOv2 runs the *k-means* algorithm on the bounding boxes of the training set to automatically find good anchor boxes.

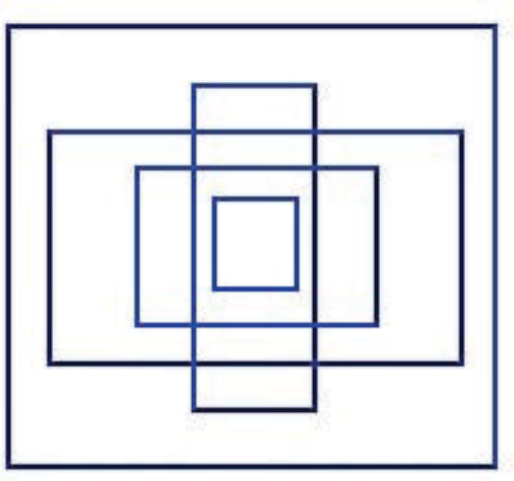


Figure 2.12: Example of anchor boxes

Another concept introduced in YOLOv2 is multi-scale training. The original YOLO uses a fixed input resolution of 448×448 pixels. In YOLOv2, every 10 batches, the network randomly chooses a new image dimension size between 320×320 and 608×608 pixels (default values). This approach forces the network to learn a significant variety of input dimensions, increasing the mAP by 1.4%.

It is necessary to point out that after YOLOv2 other versions of YOLO were proposed. However, we will not go into details since, in experiments carried out during the research, we verified that the YOLO approach met all our needs in the evaluated scenarios.

2.4 SEGMENTATION

In this section are presented the segmentation approaches employed during the development of this work. Section 2.4.1 shows the FCN approach concepts, and Section 2.4.2 explore the GAN.

2.4.1 Fully Convolutional Network-Based Semantic Segmentation

Image segmentation is a method in which a digital image is broken down into various subgroups called *image segments* which helps in reducing the complexity of the image to make further processing or analysis of the image simpler. Segmentation in easy words is assigning labels to pixels. All picture elements or pixels belonging to the same category have a common label assigned to them.

Segmentation can be achieved by using an architecture similar to the classification problem with a slight modification. Instead of one prediction on entire image, we can generate predictions for each pixel thus locating distinct classes in an image. At first, adapting an object detection architecture seems like a good idea, there are some disadvantages that have to be considered:

- A segmentation approach involves prediction at an individual pixel level, thus requiring a dense layer with an enormous number of parameters that need to be learned making it highly computationally expensive;
- the use of dense layers as final output layers leads to a constraint on the dimension of the input image. A different architecture has to be defined for different input sizes;
- No shared features are reused between overlapping patches, thus highly inefficient.

To remedy the disadvantages presented, we can stack some Convolution Layers having similar padding to preserve dimension and output a final segmentation map. This way, the model will learn the mapping from the input image to its corresponding segmentation map through the successive transformation of feature mappings.

At first stack convolution layers seems like a good idea, though there is a major issue with this Architecture as well. Preserve the input image resolution by employing the same padding in all Convolution Layers becomes quite computationally expensive.

To solve the performance problem, we could opt for a smaller number of layers, but this would greatly harm the segmentation result. We do not face this dilemma in a classification task because, for this task, we are only concerned with the presence of a single object of interest, and losing information about the location of that object is harmless. Therefore, we may periodically reduce the resolution of images through pooling. However, this is not the case with semantic segmentation.

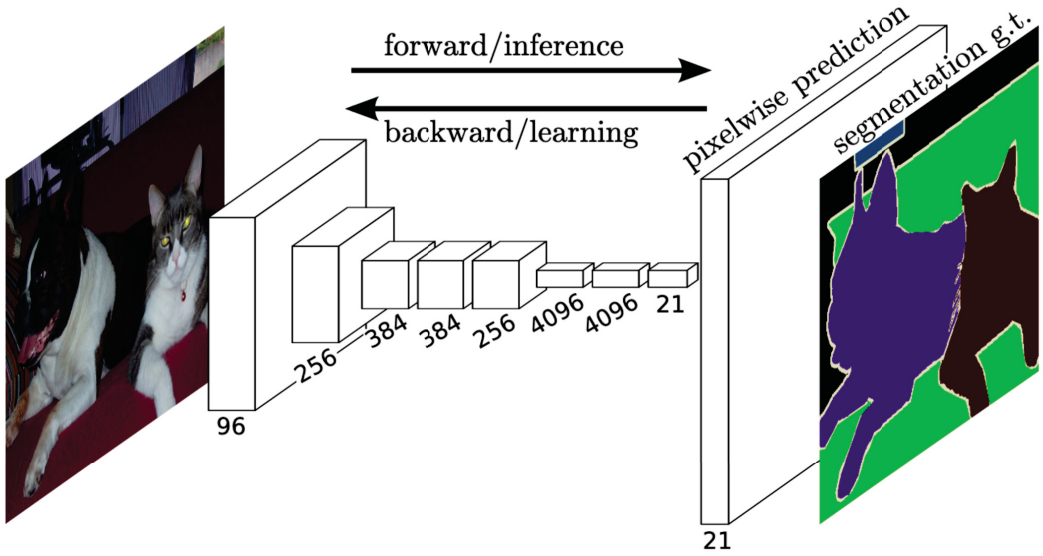


Figure 2.13: FCN skip connections overview (Long et al., 2015).

The FCN architecture proposed by Long et al. (2015) (see Figure 2.13) as one of the most popular approaches for image segmentation to solve the previous problem. In the proposed solution, the network employs the downsampling and upsampling concepts. In the first half of the model, the image is downsampled, developing complex feature mappings. With each convolution, the network captures more refined information about the image. At this stage, the network highly discriminates between different classes; however, the information about the location is lost. To recover the location information, downsampling is followed by an upsampling procedure which takes multiple lower resolution images as input and gives a high-resolution segmentation map as output

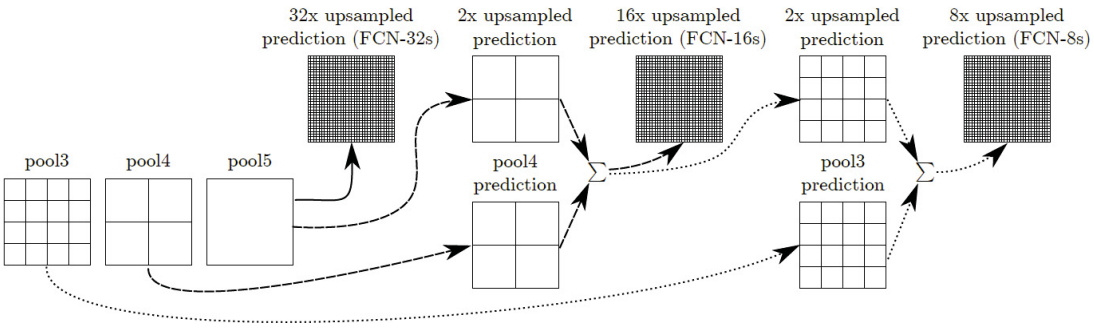


Figure 2.14: FCN skip connections overview (Long et al., 2015).

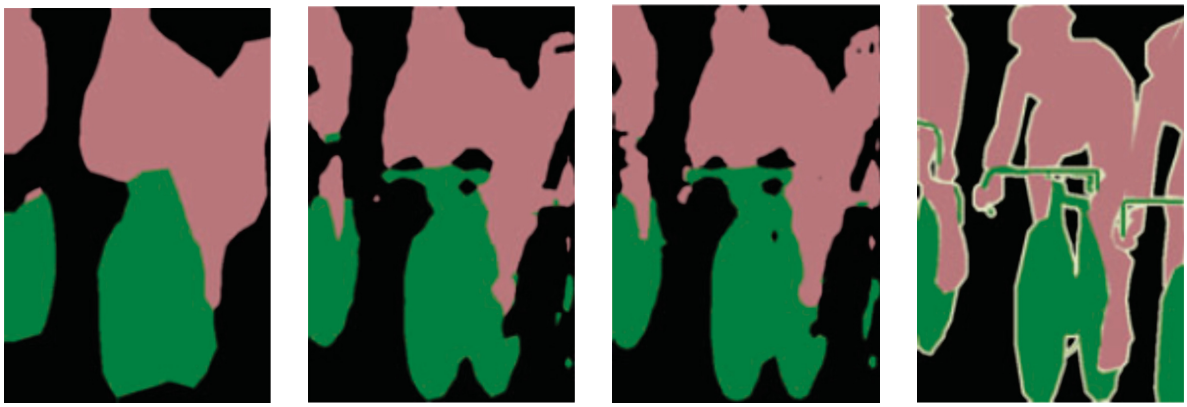
In the upsampling procedure Long et al. (2015) added *skip connections* connecting upsampling with downsampling layers. The skip connections provide enough to later layers generating accurate segmentation boundaries. This combination of initial and last layers leads to local predictions with nearly accurate global (spatial) structure. Figure illustrated a network that combine coarse layer information with fine layer information. Layers are represented as grids with relative spatial coarseness, while the intermediate convolution layers of FCN are omitted for ease in understanding.

Long et al. (2015) added skip connections connecting upsampling with downsampling layers in the upsampling procedure. The skip connections provide enough to later layers generating accurate segmentation boundaries. This combination of initial and last layers leads to

local predictions with nearly precise global (spatial) structure. Figure 2.14 illustrates a network that combines coarse and fine layer information. Layers are represented as grids with relative spatial coarseness, while the intermediate convolution layers of FCN are omitted for ease of understanding. The authors proposed three types of skip connection by combining different convolutional layers:

- **FCN-32 :** Upsamples at stride 32, predictions back to pixels in a single step (Basic layer without any skip connections);
- **FCN-16 :** Combines predictions from both the final layer and the pool4 layer with stride 16, finer details than FCN-32s;
- **FCN-8 :** Adds predictions from pool3 at stride 8, providing even further precise boundaries.

Adding Skip connections to the proposed network Long et al. (2015) presented a boosting method to CNN-based approaches since the network performance is improved by using predictions (feature maps) from previous layers. Figure 2.15 shows how accurate a segmentation result can be taking into account the proposed skipping connection approaches.



(a) FCN-32 (b) FCN-16 (c) FCN-8 (d) Ground Truth

Figure 2.15: Comparison with different FCNs. Replicated image from (Long et al., 2015).

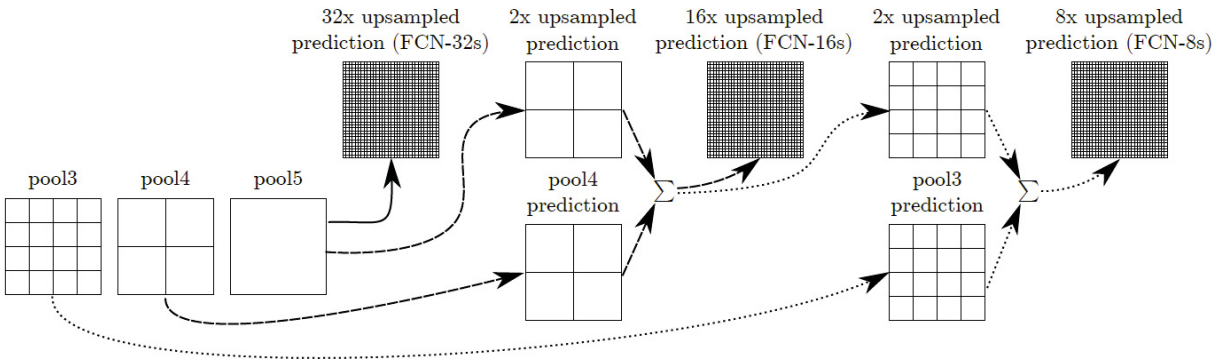


Figure 2.16: FCN architecture overview (Long et al., 2015).

2.4.2 Generative Adversarial Network

A machine learning problem involves using a computational model to make a prediction. However, to generate a computational model, a dataset comprised of multiple examples is necessary, and each one of these examples has input variables x and an output class y. The model is trained by showing examples of inputs, having it predict outputs, and correcting the model to make the output more like the expected output class y. This model correction is referred to as supervised learning (Cunningham et al., 2008).

There is another paradigm of learning where the model is only given the input variables X and the problem does not have any output variables y. A model is constructed by extracting or summarizing the patterns in the input data. There is no correction of the model, as the model is not predicting anything. This lack of correction is generally referred to as unsupervised learning. (Barlow, 1989).

The classification task is an example of supervised learning, in which we can predict a class label given a sample of input variables. Classification is traditionally referred to as discriminative modeling because a model must discriminate examples of input variables across classes (Cunningham et al., 2008).

On the other hand, an unsupervised learning approach should be able to sufficiently summarize a data distribution and then be used to generate new variables that plausibly fit into the distribution of the input variable. These types of approaches are referred to as generative models. An excellent generative model should be able to generate new examples that are not just plausible but indistinguishable from real examples from the problem domain (Barlow, 1989).

A modern example of deep learning generative model is the Generative Adversarial Network (GAN). The GAN architecture was first described in the 2014 by Goodfellow et al. (2014). However, Radford et al. (2015) have proposed a standardized approach called Deep Convolutional Generative Adversarial Networks (DCGAN), that led to more stable models.

Figure 2.17 shows the basic GAN architecture that involves two sub-models: a generator model for generating new examples and a discriminator model for classifying whether generated examples are real, from the domain, or fake, generated by the generator model.

A generator model takes a fixed-length random vector as input and generates a sample image in the domain. The input vector is drawn randomly from a Gaussian distribution and is used to seed the generative process. After training, points in this multidimensional vector space will correspond to issues in the problem domain, forming a compressed representation of the data distribution(Goodfellow et al., 2014, 2016).

This vector space is comprised of latent variables. Latent variables, or hidden variables, are those variables that are important for a domain but are not directly observable. Often refer to latent variables, or a latent space, as a projection or compression of data distribution. A latent space provides a reduction or high-level concepts of the observed raw data, such as the input data distribution. In the case of GANs, the generator model applies meaning to points in a chosen latent space. New points drawn from the latent space can be provided to the generator model as input and used to generate new and different output examples. After training, the generator model is kept and used to create new samples (Radford et al., 2015).

The discriminator is a normal (and well understood) classification model that takes input an example from the domain as (real or generated) and predicts a binary class label of real or fake (generated). The real example comes from the training dataset and the generated examples are the output of the generator model. Commonly after the training process, the discriminator model is discarded as we are interested in the generator (Goodfellow et al., 2014).

However, sometimes the generator can be repurposed as it has learned to extract features from examples in the problem domain effectively. Some or all of the feature extraction layers can

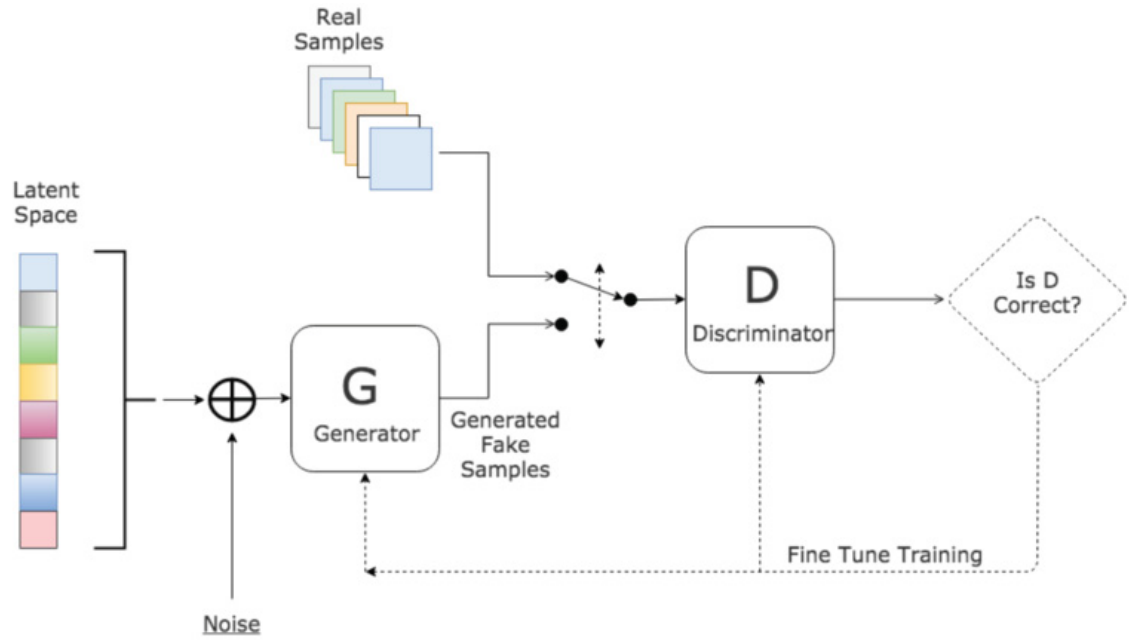


Figure 2.17: GAN architecture overview. Replicated image from https://towardsdatascience.com/semi-supervised-learning-and-gans-f23bbf4ac683.

be used in transfer learning applications using the same or similar input data (Goodfellow et al., 2016).

Generative modeling is an unsupervised learning problem, as we discussed previously. However, a clever property of the GAN architecture is that the training of the generative model is framed as a supervised learning problem. Both generator and discriminator are trained together. The generator generates a batch of samples, and these, along with real examples from the domain, are provided to the discriminator that classifies these as real or fake (Goodfellow et al., 2016). The discriminator is then updated to get better at discriminating real and fake samples in the next round, and importantly, the generator is updated based on how well, or not, the generated samples fooled the discriminator (Goodfellow et al., 2014).

In a simplified way we can say that, the two models are competing against each other, they are adversarial in the game theory sense, and are playing a zero-sum game. In this case, zero-sum means that when the discriminator successfully identifies real and fake samples, it is rewarded or no change is needed to the model parameters, whereas the generator is penalized with large updates to model parameters. Alternately, when the generator fools the discriminator, it is rewarded, or no change is needed to the model parameters, but the discriminator is penalized and its model parameters are updated (Radford et al., 2015).

2.5 FINAL REMARKS

In this chapter are presented the most relevant detection and segmentation approaches employed on detection and segmentation scenarios. In next chapter are described the literature review. First, some works on iris and sclera detection/segmentation are presented, and in the next section RoI extraction by using contextual relationships concepts are presented.

3 LITERATURE REVIEW

In this chapter, we briefly review the most relevant approaches employed to extract the Ocular Region Components (ORC). More specifically, in Section 3.1.1 we investigate the most relevant detection approaches, and in Section 3.1.2, we evaluate the segmentation approaches.

In addition to investigating the ORC extraction approaches in Section 3.2 we investigate the works in which the contextual relationship among the elements present in an image were explored. In all the previously described sections the works are presented in chronological order aiming to illustrate the evolutionary process of the research area.

3.1 REGION OF INTEREST EXTRACTION

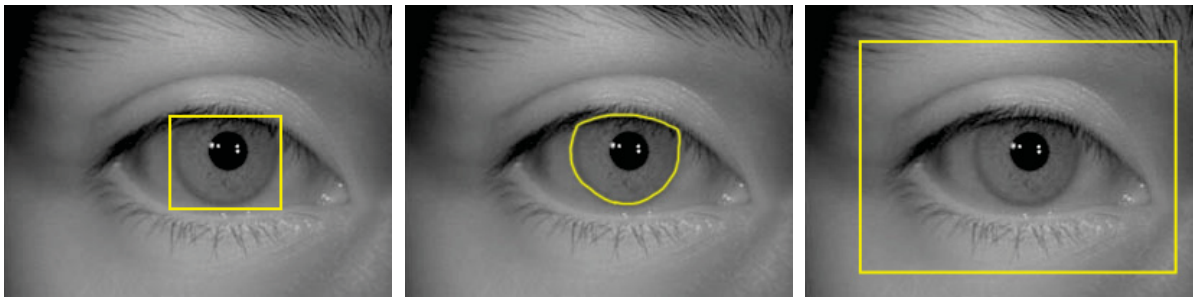
In this section the most commonly used approaches in the ORC extraction are presented. In this way, Section 3.1.1 and Section 3.1.2 describes the most relevant works in detection and segmentation of the ORC respectively.

3.1.1 Detection

This section presents the main works from the literature that aims to solve the problem of detection of the components of the ocular region. Table 3.1 summarizes all the works described in this section, , showing the detection approach, segmented region, database, and results obtained for each studied method.

3.1.1.1 Iris Detection

Given the importance of the preprocessing step in the periocular region based biometrics, many authors exerted efforts to solve this problem by proposing different ways to represent the Region of Interest (RoI) (Daugman, 1993; Wildes, 1997; Zhou et al., 2013; Zhang and Ma, 2014; Su et al., 2017; Severo et al., 2018; Park et al., 2011; Le et al., 2014; Proença et al., 2014).



(a) (b) (c)

Figure 3.1: RoI Extraction: 3.1(a) Rectangular iris bounding box; 3.1(b) Elliptical outer iris contour sample; 3.1(c) Rectangular periocular region bounding box sample.

Table 3.1: Eye region detecion approaches

Author(s)	Detection Approach	Region	Database	Metrics	Best Result %
Daugman (1993)	Integro-Differential Operator	Iris	Own Database	N/A	N/A
Wildes (1997)	Integro-Differential Operator and the Hough-Transform	Iris	Own Database	N/A	N/A
Rodríguez and Rubio (2005)	Integro-Differential Operator	Iris	CASIA-IrisV1	Accuracy	98.00
Zhou et al. (2013)	Integro-Differential and Vector Field Convolution (VFC)	Iris	CASIA-IrisV2	Accuracy	98.85
Zhang and Ma (2014)	Integro-Differential and momentum-based level	Iris	CASIA-IrisV2	Accuracy	98.53
Alvarez-Betancourt and Garcia-Silvente (2010)	QMA-OWA	Iris	CASIA-IrisV3	Accuracy	98.00
Cui et al. (2012)	Dual-threshold and Hough-Transform	Iris	CASIA-IrisV3-Twins	Accuracy	98.00
Su et al. (2017)	Iterative searching	Iris	CASIA-IrisV1 and CASIA-IrisV3	Accuracy	98.08
Severo et al. (2018)	Convolutional Neural Network (CNN)	Iris	IIIT-D CLI NDCLD15 MobBIOfake NDCCL CASIA-IrisV3 Interval BERC mobile-iris database	Accuracy	98.33 98.54 98.90 99.05 97.10 99.71
(Park et al., 2011)	Anthropometry of the human face	Periocular	FRGC	N/A	N/A
(Juefei-Xu and Savvides, 2012)	Active Shape Model (ASM)	Periocular	Private	N/A	N/A
(Mahalingam et al., 2014)	Average of Synthetic Exact Filters (ASEF)	Periocular	Private	N/A	N/A
(Le et al., 2014)	Local Eyebrow Active Shape Model (LE-ASM)	Periocular	AR Face Dataset and MBGC	N/A	N/A
(Proença et al., 2014)	Markov Random Field (MRF)	Periocular	UBIRIS.v2 subset	N/A	N/A

Figure 3.1 shows the most consistent RoI extraction approaches presented in the literature. In the iris detection scenario, two different representations are used to delimit the iris, one based on a rectangular region around the RoI, and another based on the elliptical outer boundary. In turn, the periocular region is delimited just by a rectangular form around the its boundaries.

Many works present in the literature show the iris delimitation by using an elliptical contour around the outer edge of it. Daugman (1993) was a pioneer in this scenario, proposing an approach that makes the use of an integro-differential operator to detect the iris, identifying the borders present in the images. This operator takes into account the circular shape of the iris to found the correct position of it, by maximizing the partial derivative concerning the radius. The author employed an own database to perform the experiments, being this composed of 592 eye images from 323 subjects, captured in the Near-Infrared (NIR) wavelength. However, the results about the detection of iris were not be informed since the focus of the work was people identification by using the iris.

Similar to Daugman (1993), Wildes (1997) also has proposed another proper method for iris location, and to show the robustness of the proposed approach the author applies it to a new d database composed of 600 images from 40 subjects. In the new presented method, the iris is detected using border detection and the Hough-Transform. To perform the location, the iris is isolated by using Gaussian filters of low pass followed by a spatial sub-sampling. Subsequently, a Hough transform is applied, and those elements that better fit a circle according to a defined condition are selected. The results of the proposed approach were not reported since the focus of the work is iris recognition, and the detection was employed as a preprocessing step.

Authors such as Rodríguez and Rubio (2005) and Zhang and Ma (2014) also employes the integro-differential operator proposed by Daugman to perform the iris location. In the approach proposed by Rodríguez and Rubio (2005), a two-stage method based on the site of inner and outer iris contours was presented. In the first step, the iris inner contour is located by using Daugman's operator. Then, the location of the outer boundary is determined by detecting the triangle inscribed in the iris circumference. The presented approach, achieves 98% of accuracy on CASIA-IrisV1 database (CASIA, 2002).

In work presented by Zhang and Ma (2014), the authors adopted a method that uses momentum-based level set (Läthén et al., 2009; Wang et al., 2010) together with the Daugman's operator to locate the pupil boundary. In the proposed method initially the contour of iris is obtained with a momentum-based level set using the minimum average gray level, and after that the integro-differential operator is applied to make the detection, decreasing the time consumption. By using the new approach the authors achieves an accuracy of 98.53% on CASIA-IrisV2 database (CASIA, 2004). This improvement happens because the initially detected contour is generally close to the real iris inner boundary.

Aside from the integro-differential based approaches, many other works were proposed to the extraction of the periocular region components such as Alvarez-Betancourt and Garcia-Silvente (2010), Cui et al. (2012), Zhou et al. (2013), Su et al. (2017) and Severo et al. (2018).

Alvarez-Betancourt and Garcia-Silvente (2010) presented an iris location method based on the detection of circular boundaries using an approach of gradient analysis in points of interest of successive arcs, and achieves an accuracy of 98% on CASIA-IrisV3 database (CASIA, 2010), with improvements in processing time. To perform the iris location the quantified majority operator QMA-OWA proposed by Peláez and Doña (2006) was used to obtain a representative value for each successive arc, and the iris boundary is given by getting the arch with the most significant representative amount.

In the method proposed by Cui et al. (2012), the eyelashes are removed as a first step using the dual-threshold method, which can be an advantage over other iris location approaches.

Next, the facula is removed by using erosion mathematical morphology. Finally, the accurate iris position is obtained through Hough-Transform and least-squares. The proposed method achieved 98% accuracy in CASIA-IrisV3-Tbiwins database (CASIA, 2010) showing a strong anti-interference ability, detecting the iris inner and outer edges effectively.

Zhou et al. (2013) have presented a method for iris location in which the initial position of the this is obtained by using the Vector Field Convolution (VFC) technique. This initial estimation makes the location of the pupil much easy and closer to the real boundary instead of circle fitting, improving location accuracy, and reducing computational cost. The final result is obtained by using the Daugman's approach, and the main contribution of the proposed method is a shorter computational cost improving the location accuracy, since the pupil delineation is much closer to the real boundary, achieving 98.85% on CASIA-IrisV2 database (CASIA, 2004).

Su et al. (2017) proposed an iris location algorithm based on local property and iterative searching, achieving 98.08% accuracy on CASIA-IrisV1 and CASIA-IrisV3 databases. To detect the RoI the pupil area is extracted using iris regional attributes, and the inner edge of it is fitted by iterating, comparing and sorting the pupil edge points. The outer edge location is made by using an iterative searching method from the extracted pupil center and radius, with a shorter time in relation to the approaches available in the literature.

Severo et al. (2018) have proposed a Convolutional Neural Network (CNN) based approach to represent the iris location by using a rectangular bounding box. To perform the Region of Interest (RoI) extraction, the You Only Look Once (YOLO), a well-known deep learning approach in detection scenario, is employed by using a fine-tuned model, overcoming problems such as noise, eyelids, eyelashes, and reflections. Six databases were used for the experiments performed: IIIT-Delhi Contact Lens Iris (IIIT-D CLI) (Kohli et al., 2013), Notre Dame Contact Lens Detection 2015 (NDCLD15) (Doyle and Bowyer, 2015a), MobBIO (Sequeira et al., 2014b), Notre Dame Cosmetic Contact Lenses (NDCCL) (Doyle and Bowyer, 2014), CASIA-IrisV3-Interval (CASIA, 2010) and BERC mobile-iris database (Kim et al., 2016), and their respective accuracies are: 98.33%, 98.54%, 98.90%, 99.05%, 97.10%, and 99.71%.

3.1.1.2 Periocular Detection

Not only works related to iris detection can be found in the literature. On the periocular region detection scenario, some works are presented, however, in scenario the evaluation metrics are not reported once the RoI detection is only a preprocessing step used in biometric systems.

Park et al. (2011) have proposed one of the first periocular region based biometric approaches, and to achieve the main aim objective of the work, the authors have presented an eye region detector. The proposed eye detector take as input face images detected by using the Viola-Jones detector (Viola and Jones, 2001) applied in the Face Recognition Grand Challenge (FRGC) database, and output the periocular area.

Such as Park et al. (2011), Juefei-Xu and Savvides (2012) have proposed a periocular region detection approach that takes as input a face image detected by using the Viola-Jones detector. However, the proposed approach were applied on a private database composed by 3,200 of size images from 40 subjects, and the periocular region is identified by using a Active Shape Model (ASM) approach that identifies 79 facial landmarks, and among these are points respective to eye region.

As well as in the early presented works Mahalingam et al. (2014) have proposed a eye detector method that needs of a face image as input. The presented detector receives a face image previously obtained rectangular face image and output the periocular region by using the proposed by Bolme et al. (2009). All the performed experiments was made on a private database composed by 1.2 million faces from 38 subjects.

Le et al. (2014) have proposed the Local Eyebrow Active Shape Model (LE-ASM) to detect the eyebrow region directly from a given face image, and after detect the periocular region by using the output of the new ASM approach. The author have evaluated the periocular region result in the AR Face Dataset and in the mbgc databases, however, as well as in all the works presented in this section at this moment the evaluation results were not presented.

Proença et al. (2014) proposed a Markov Random Field (MRF) method to segment the periocular region components, and another elements around this region (iris, sclera, eyelashes, eyebrows, hair, skin and glasses). The presented MRF approach analyzes the image pixels and output the segmented region taking into account the appearance and geometrical constraints assuring that the output of the system is biologically plausible. Although the objective of the presented technique is segment the periocular region elements and the region around this, the periocular region can be predicted by combining the outer limits of the sclera and the lower eyelashes.

3.1.2 Segmentation

As it was previously presented, the segmentation, as well as the detection, have great importance in the biometric systems. This importance is due to the fact of an incorrect extraction of the RoI can affect the effectiveness of the whole system.

Taking into account the segmentation importance, Daugman (1993) proposed the determination of an annular region around the iris, and this region is considered as segmentation. However, many other methods were proposed from those that use conventional approaches of image processing to deep learning approaches. Table 3.2 summarizes the works presented in this section, showing the segmentation approach, segmented region, database, and results obtained for each studied method.

3.1.2.1 Iris Segmentation

From the techniques that use a conventional approach of image processing, deserve to be highlighted the works presented by Liu et al. (2005), Proença and Alexandre (2006), Shah and Ross (2009), Tan et al. (2010), Podder et al. (2015), Haindl and Krupička (2015) and Ouabida et al. (2017).

Liu et al. (2005) have presented a two stages approach based on Daugman's integro-differential method. In the first step, the internal limit of the iris is detected to later identify its external perimeter. After the perimeter delimitationr, the noise pixels based on the high/low-intensity level are eliminated. The proposed method reaches 97.08% accuracy on its dataset composed of 4000 images.

Proença and Alexandre (2006) used a different approach from the early presented works where each pixel of the image is classified considering its coordinates and density distribution by using the Fuzzy K-means algorithm. Then the Canny edge detector is applied on the image creating an edge map, and finally, the circular Hough-Transform detects the inner and outer iris boundaries, with 98.02% accuracy on UBIRIS V1 database (Proença and Alexandre, 2005).

Shah and Ross (2009) performed the iris segmentation through Geodesic Active Contours, combining energy minimization with active contours based on curve evolution. The pupil is detected by using a binarization technique, and both inner and outer iris boundaries are approximated using the Fourier series coefficients. The proposed method achieves 94% accuracy on CASIA-IrisV3-Interval (CASIA, 2010) database.

The approach presented by Tan et al. (2010), the winner of Noisy Iris Challenge Evaluation, Part I (NICE-I), achievs 100% accuracy on UBIRIS v1 database by using an approach

Table 3.2: Ocular region segmentation approaches

Author(s)	Detection Approach	Region	Database	Metrics	Best Result %
Liu et al. (2005)	Integro-Differential Operator	Iris	Own Database	Accuracy ER	97.08 01.79
Proença and Alexandre (2006)	Fuzzy K-Means	Iris	UBIRIS v1	Accuracy	98.02
Shah and Ross (2009)	Geodesic Active Contours (GAC)	Iris	CASIA-IrisV3-Interval	Accuracy	98.00
Tan et al. (2010)	Integro-Differential and Clustering	Iris	UBIRIS v1	Accuracy	100.00
Podder et al. (2015)	Radial Suppression	Iris	CASIA-IrisV3-Interval CASIA-IrisV3-Lamp	ER	00.54 00.82
Haindl and Krupička (2015)	Multi Spatial Probability and Adaptative Threshold	Iris	UBIRIS v2	ER	00.01
Ouabida et al. (2017)	Vander Lugt Correlator and Active Contours	Iris	CASIA-IrisV4-Interval UBIRIS v2	ER	01.44 00.80
Liu et al. (2016a)	Hierarchical Convolutional Neural Network (HCNN) Multi-scale Fully Convolutional Network (MFCN)	Iris	CASIA-IrisV4-Distance UBIRIS V2	ER	00.59 00.90
Bezerra et al. (2018)	CNN	Iris	BioSec CASIA-IrisV3-Interval CASIA-IrisV4-Thousand IIIT-D CLI NICE.I CROSS-EYED MICHE-I	ER	00.58 00.55 01.25 00.72 02.67 01.12 01.90
Osorio-Roig et al. (2018)	Fully Convolutional Network (FCN)	Iris	NICE.I MobBIO	ER	1.13 2.46
Zhao and Kumar (2019)	UniNet.v2	Iris	ND-Iris-040 CASIA-Iris-Distance IITD	ER	1.68 0.67 5.34
Wang et al. (2020a)	IrisParseNet	Iris	CASIA-IrisV4-Distance, the UBIRIS.v2 and the	$F-Score$	94.25 91.78 and 93.05%
Mashudi et al. (2021)	Dynamic-U-Net	Iris	Mobile Iris Challenge Evaluation I (MICHE-I)	$F-Score$	94.59
Radu et al. (2015)	Statistical Image Features, Zernike Moments and Histogram of Oriented Gradients (HOG)	Sclera	SSBC 2015 Database	Precision Recall	95.05 94.56
Rot et al. (2018)	SegNet	Iris Sclera Pupil Periocular Eyelashes Canthus	MASD	F-Score	91.00 91.00 85.00 90.00 63.00 49.00
Lucio et al. (2018)	FCN	Sclera	UBIRIS V2 MICHE-I MICHE-GS4 MICHE-IP5 MICHE-GT2	F-Score	87.48 88.32 88.12 87.80 87.94
Naqvi and Loh (2019)	Sclera-Net	Sclera	MICHE-GS4 MICHE-IP5 MICHE-GT2 SBVPI	F-Score	93.49 92.66 93.26 92.66
Wang et al. (2020b)	ScleraSegnet	Sclera	UBIRIS V2 MICHE-I MICHE-GS4 MICHE-IP5 MICHE-GT2	F-Score	91.43 89.54 90.45 89.28 89.34
Vitek et al. (2021)	UPerNet U-Net	Sclera	SBVPI V2 MASD-I	F-Score	95.50 92.70

that removes the reflection points present in the image. These points are removed by using adaptive threshold and bi-linear interpolation, combined with a growing region technique that uses clustering and Integro-differential operator.

Already Podder et al. (2015) applied a Maximum Radial Suppression (MRS) technique to noise removal together with the Canny edge detector and Hough-Transform to detect iris

boundaries Using the proposed technique over CASIA-IrisV3-Interval and CASIA-IrisV3-Lamp (CASIA, 2010) databases the respective Error Rate (ER) are achieved 0.54% and 0.82%.

Haindl and Krupička (2015) have presented an approach in which the eyelids and the specular highlights are removed from the image, to after segment the iris. The third-order polynomial mean combined with the standard deviation were employed to remove the eyelids, while the specular highlights are removed by using the adaptive thresholding and Markovian Texture Model (MTM). Using the proposed technique, a 0.01 EER was obtained on UBIRIS v2 (Proenca et al., 2010) database. While Ouabida et al. (2017) achieves 1.44% and 0.80% ERs on CASIA-IrisV4-Interval and UBIRIS v2 databases, employing an approach based on the Optical Correlation based Active Contours (OCAC) to detect the iris and pupil contours through spatial filtering.

With the advent of deep Convolutional Neural Networks (CNNs) provided by the immense computational power offered by the Graphical Processor Units (GPUs) many computer vision problems were solved (Ahuja et al., 2017; Dumoulin and Visin, 2016; Severo et al., 2018). In this way, Teichmann et al. (2016a) have proposed a CNN architecture, called MultiNet, to merge detection, classification, and semantic segmentation. Inspired by the excellent results reported by Teichmann et al. (2016a), Bezerra et al. (2018) have proposed an iris segmentation approach based on the decoder module of the MultiNet. By using the proposed method, the authors achieved state-of-the-art results in the employed databases. The databases employed were: BioSec (Fierrez et al., 2007), CASIA-IrisV3-Interval, CASIA-IrisV4-Thousand, IIIT-D CLI, NICE.I (Proença and Alexandre, 2012), CROSS-EYED (Sequeira et al., 2016) and MICHE-I (Marsico et al., 2015), and they respective ERs are: 0.58%, 0.55%, 1.25%, 00.72%, 2.67%, 1.12%, and 1.90%.

Liu et al. (2016a) also have explored the deep learning world in iris segmentation proposing two new approaches called Hierarchical Convolutional Neural Network (HCNN) and Multi-scale Fully Convolutional Network (MFCN) to perform a dense prediction of the pixels using sliding windows, merging shallow and deep layers, achieving 0.90% and 0.59% ERs over UBIRIS V2 and CASIA-IrisV4-Distance (CASIA, 2010) databases.

Osorio-Roig et al. (2018), similarly to Bezerra et al. (2018), have also presented a FCN approach to iris segmentation. In the proposed approach, the authors employed a multi-class training, however, only the iris segmentation was evaluated. To evaluate the proposed approach, the NICE-I competition protocol was employed in two well-know iris datasets, MobBIO (Sequeira et al., 2014a) and NICE-I, obtaining ERs of 2.46% and 1.13%, respectively.

Zhao and Kumar (2019) proposed the UniNet.v2, a new iris segmentation approach, and just like Osorio-Roig et al. (2018) used the NICE-I competition protocol to evaluate the obtained results. The datasets employed to evaluate the UniNet.v2 are: ND-Iris-040, CASIA-Iris-Distance and IITD, obtaining ERs of 1.68%, 0.67%, and 5.34%, respectively.

Wang et al. (2020a) proposed the IrisParseNet, a U-Net based multi-task segmentation approach that can simultaneously predict the iris inner and outer boundaries, by using a new attention module presented by the authors. To evaluate the performance of the IrisParseNet the authors employ the CASIA-IrisV4-Distance, the UBIRIS.v2 and the datasets, achieving the $F-Scores$ of 94.25%, 91.78%, and 93.05%, respectively.

Mashudi et al. (2021) have presented a Dynamic-U-Net based iris segmentation, in the proposed approach the authors employ the ResNet-34 as feature map extractor. To evaluate the proposed approach the authors employd the Mobile Iris Challenge Evaluation I (MICHE-I) dataset achieving a $F-Score$ value of 94.59%

3.1.2.2 Sclera Segmentation

Despite the consolidation of the iris as a biometric resource, other elements that compound the periocular region have started to arouse the interest of the researchers. The sclera besides the iris was one of the first periocular region components to stimulate the attention of the researchers, once the biometrics-based in this achieves acceptable results (Delna et al., 2016). Taking in mind the good results on sclera based biometrics and the need of a proper delimitation of the ROI, as well as the in iris-based biometrics, Das et al. (2015) have proposed the first challenge of sclera segmentation in which several teams participated, and the best result was obtained by Radu et al. (2015).

Radu et al. (2015) make the use of a pixel-level algorithm, that explores the image color space with two stages classifier. In the first stage, a set of simple classifiers are employed to generate space probabilities, and in the second and final stage a neural network operates over the previously obtained probabilities, and the decision is presented using precision and recall metrics. The proposed technique achieves 95.05% and 94.56%, precision and recall values respectively on the SSBC 2015 Database.

After the good results present in the first sclera segmentation competition, other challenges were proposed (Das et al., 2016, 2017), and year after year the results presented by the teams have outperformed the results of the last year. In the benchmark proposed by Das et al. (2016) the best extraction approach is achieved using a method based on Fuzzy C Means, which considers spatial information and uses Gaussian kernel function to calculate the distance between the center of the cluster and the data points, achieving 85.21% and 80.21% precision and recall values respectively. After, in the competition proposed by Das et al. (2017) the best result achieves 95.34% and 96.65% precision and recall values respectively by using for the first time a deep learning approach to sclera segmentation.

Based on the growing interest in sclera segmentation and its extreme importance in non-invasive biometric systems, some authors identified the need to assess the feasibility of exploring the use Convolutional Neural Network (CNN) in this scenario. In this way, Lucio et al. (2018), Rot et al. (2018) and Wang et al. (2019) have presented new approaches to sclera segmentation.

Lucio et al. (2018) have proposed two new sclera segmentation approaches. The first of these based on the Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), and second based on Fully Convolutional Networks (FCNs) (Long et al., 2015). The best results were obtained by using the FCN approach achieving 87.48%, 88.32%, 88.12%, 87.80% and 87.94% F-Score values on UBIRIS V2, MICHE-I, MICHE-GS4, MICHE-IP5, and MICHE-GT2 databases, respectively.

Rot et al. (2018) proposed a multi-class segmentation approach based in the SegNet to segment simultaneously the sclera and another components from the periocular region (iris, pupil, periocular, eyelashes and canthus). The authors employed the Multi-Angle Sclera Database (MASD) in the evaluation of the proposed approach, obtaining F-Score vales of 91.00%, 91.00%, 85.00%, 90.00%, 63.00% and 53.00% on the iris, sclera, pupil, periocular, eyelashes and canthus regions, respectively.

Naqvi and Loh (2019) proposed the Sclera-Net, a residual encoder decoder network, that exploits identity and not identity mapping residual skipping connections to take benefit of the high frequency information from the prior layers of booth encoder and decoder networks to determine the accurate sclera region. To evaluate the proposed segmentation approach the authors employed the MICHE-GS4, MICHE-IP5, and MICHE-GT2 and the Sclera Blood Vessels, Periocular and Iris (SBVPI) datasets achieving 93.49%, 92.66%, 93.26%, 92.66% $F-Score$ values respectivelly.

Wang et al. (2020b) presented a new sclera segmentation approach based on attention assisted model named ScleraSegnet. The proposed method user a U-Net with some improvements in the attention modules,by doing the improvements the authors help the model to implicitly learn how to suppress irrelevant regions in an input image, while highlighting salient features useful for a specific task. By using the the ScleraSegNet the authors achieved $F - Score$ values of 91.43%, 89.54%, 90.45%, 89.28% and 89.34% on the UBIRIS V2, MICHE-I, MICHE-GS4, MICHE-IP5, and MICHE-GT2 databases, respectively.

Vitek et al. (2021) have proposed 5 deep neural network architectures to sclera segmentation being these: Segnet; DepLabv3+; HRNetV2; UperNet; and U-Net. For each one of the proposed approaches the authors investigated the results in 2 images dataset (SBVPI and MASD). In the SBVPI dataset the UperNet achieved the best $F - Score$ (95.55%), while in the MASD dataset the highest achieved $F - Score$ was obtained by using the U-Net (92.70%).

3.2 REGION OF INTEREST EXTRACTION USING CONTEXTUAL RELATIONSHIP

In addition to the commonly employed approaches to extract the eye region components presented in the previous sections, there is a set of techniques which aims to detect/segment elements on an image based on the contextual information in which is located. The contextual information can be defined as any information that is not directly produced by the appearance of an object, such as, nearby data, image tags, annotations or the presence of other objects (Kumar and Hebert, 2005; Hoiem et al., 2008; Tu and Bai, 2010). Table 3.3 shows a compilation of the works presented in this section.

Table 3.3: Context based region of interest extraction approaches.

Author(s)	RoI Extraction Approach	Context Type	Object of Interest	Database	Metric	Result %
Strat and Fischler (1991)	Camera's meta-data	Scale	Scene Understanding	Own Database	N/A	N/A
Bar and Ullman (1996)	Local features	Spatial	Object Recognition	Drawings Database	Mean Error Rate (MER)	26.00
Torralba (2003)	Correlation between the statistics of low-level features	Semantic	Scene Understanding	N/A	N/A	N/A
Kumar and Hebert (2005)	Pairwise Relationship	Spatial	Scene Understanding	Beach Dataset Sowerby Dataset Building/Road/Car dataset Monitor/Keyboard/Mouse Dataset	Accuracy	74.00 89.30 84.36 Quantitative result N/A
Wolf and Bileschi (2006)	Semantic Layers	Semantic	Scene Understanding	StreetScenes Database	Accuracy	83.00
Rabinovich et al. (2007)	Conditional Random Fields (CRFs)	Semantic	Scene Understanding	Pascal VOC 2007 MSRC	Accuracy	74.20 68.40
Verbeek and Triggs (2007)	Conditional Random Fields (CRFs)	Semantic	Scene Understanding	MSRC Sowerby Dataset Corel Dataset	Accuracy	84.90 87.40 74.60
Galleguillos et al. (2008)	Co-occurrence and relative location contexts	Semantic	Scene Understanding	Pascal VOC 2007 MSRC	Accuracy	36.70 68.47
Shotton et al. (2009)	Inter pixels statics	Scale	Scene Understanding	MSRC Corel Database Subset Sowerby Database Subset Set of video sequences of television shows	Accuracy	72.20 74.60 88.60 68.88
Leng et al. (2018a)	Context Learning Network (CLN)	Scale	Scene Understanding	Pascal VOC 2007 PASCAL VOC 2012 Coco Database	mean Average Precision (mAP)	82.10 80.70 38.40
Leng et al. (2018b)	Context-aware U-Net	Scale	Iris	ISBI Challenge Database	Warping Error Rand Error Pixel Error	0.000121 0.021200 0.034600
Zhang et al. (2020)	Contextual Bidirectional Enhancement (CBD-E)	Spatiak	Object Detection	NWPU RSOD DIOR	mAP	94.98 94.23 67.80
Cai et al. (2021)	Cross-Channel Feature Pyramid Network (CFPN) combined with Foreground Attentive Detection Heads (FDH)	Scale	Object Detection	NWPU RSOD DIOR	mAP	96.98 95.60 73.50.80
Liu et al. (2022)	Balanced Feature Pyramid Network (BFPN) combined with Task-interactive Head (TIH)	Spatial Semantic	Object Detection	HRRSD DIOR	mAP	69.49 85.20 73.50

A traditional object categorization approach is made using appearance features such as color, edges responses, texture, and shapes cues. However, according to Biederman (1972), using only these features set is not possible to understand a scene composition. To detect an object based on context is necessary to understand the five different classes of relationships between an object and its surroundings: *interposition*, *support*, *probability*, *position*, and *size*. *Interposition* and *support* refer to the physical space. *Probability*, *position*, and *size* are defined as semantic relations once they require access to the referential meaning of the object. Semantic relations include information about specific interactions among objects in the scene, and they are often used as *contextual features*.

Many authors have explored the semantic relationships proposed by Biederman (1972) with the purpose of improving the results on objects recognition task (Torralba, 2003; Fink and Perona, 2003; Carbonetto et al., 2004; He et al., 2004; Rabinovich et al., 2007; Galleguillos et al., 2008). These relationships can be grouped into three categories: semantic context (*probability*), spatial context (*position*), and scale context (*size*).

Semantic context corresponds to the likelihood of an object to be found in some scenes but not in others. Hence, it is defined in terms of the co-occurrence of one object with others and its occurrence in scenes. Trying to solve this problem Fischler and Elschlager (1973), Hanson (1978) and Strat and Fischler (1991) have proposed approaches based on predefined rules.

After more than a decade since the presentation of the last significant work taking into account the semantic context, Torralba (2003) presented one of the first methods that makes use of statistical approaches to detect objects. The proposed technique explores and generalizes the semantic context in the real world, by using the correlation among the statistics of low-level features across the entire scene and the objects that it contains. However, the authors do not have reported the results obtained, and the database employed in this work.

In the same direction of Torralba's work, other authors such as Wolf and Bileschi (2006); Rabinovich et al. (2007); Verbeek and Triggs (2007); Galleguillos et al. (2008) have employed statistical methods to detect objects based on semantic context.

Wolf and Bileschi (2006) have proposed an approach to extract the semantic context by using context features and the Support Vector Machines (SVM) classifier. The presented approach achieved an accuracy of 83.00% (estimated value obtained from a Receiver Operating Characteristics (ROC) curve) on their database named StreetScene. To obtain the context features the authors proposed a two stages process. In the first stage, the image is processed to calculate the low level and semantic information. In the second stage, the context feature is calculated at each point by collecting samples of the previously computed features at predefined relative positions.

Rabinovich et al. (2007) have employed the Conditional Random Field (CRF) to detect the objects in a scene maximizing the labels agreement according to the contextual relevance. Firstly, a fully connected graph among the segmented labels was used. An then, the CRF was trained in more straightforward problems defined on a relatively small number of segments. The proposed approach showed to be competitive with the state-of-the-art achieving 74.20% and 68.40% accuracies on Pascal VOC 2007 (Everingham et al., 2007) and Microsoft Research Cambridge (MSRC) (Ali and Zafar, 2018) databases respectively.

Verbeek and Triggs (2007) have proposed a segmentation approach that achieves 84.90%, 87.40% and 74.60% accuracies on the MSRC, Sowerby (He et al., 2004) and Corel[1] databases respectively. The previously mentioned results were obtained by using CRF together with unlabeled nodes. In this way the unknown labels are marginalized and the log-likehood of the known labels can be maximized by gradient descendent.

[1]https://sites.google.com/site/dctresearch/Home/content-based-image-retrieval

Galleguillos et al. (2008) introduced a new approach to object categorization based on two types of context (co-occurrence and relative location) by using local appearance features and achieved 36.70% and 68.47% accuracies, on the Pascal Voc 2007 and MSRC databases. The presented method uses CRFs to maximize object label attribution to both semantic and spatial features. The results of this work showed that combining co-occurrence and spatial context improves the accuracy of the system when compared with the results where only co-occurrence data are used in training.

In the spatial context scenario Bar and Ullman (1996) explored the consequences of spatial relations on human performance in recognition tasks, and reported a 26.00% Mean Error Rate (MER) on a drawings database. The results presented in this work suggest that: (i) the presence of objects that have a unique interpretation improve the recognition of ambiguous objects; and (ii) proper spatial relationships among objects decreases error rates in recognition of individual objects. The affirmations mentioned above refer to the use of semantic and spatial context to identify ambiguous objects, once the spatial context implicitly encodes the co-occurrence of other objects in the scene and offers more specific information about the configuration in which those objects are usually found.

Taking as a starting point the work presented by Bar and Ullman (1996), Shotton et al. (2009) proposed an approach that explores the inter-pixels statistics to discriminates objects by using texture, layout, and spatial information. The discrimination among the object is made capturing the features mentioned above and incorporating these in a CRF. The experiments achieve 72.20%, 74.60%, 88.60%, and 68.88%, on MSRC database, Corel Database subset, Sowerby Database subset and a set of video sequences of television shows respectively.

Still, taking into account the spatial context Kumar and Hebert (2005) explored a pairwise relationship by using a two-layer hierarchical formulation to exploit different levels of contextual information in images. In the first layer, it encodes the region's interaction, and in the second, the object's interactions are mapped, as shown in Figure 3.2. The experiments were performed on four databases: Beach Database (Kumar et al., 2003), Sowerby Database (He et al., 2004), Building/Road/Car dataset (Torralba et al., 2004) and Monitor/Keyboard/Mouse Dataset (Torralba et al., 2004), and the respective accuracies for these are 74.00%, 89.30%, 84.36% and 90.00% (result obtained analyzing a Receiver Operating Characteristics (ROC) curve).

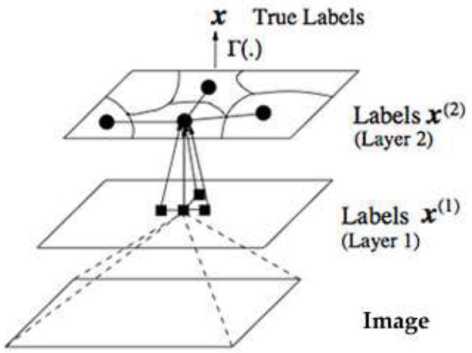


Figure 3.2: Sample of the architecture proposed by (Kumar and Hebert, 2005).

As was previously presented, the semantic context is implicitly present in the spatial context, since the information of object co-occurrences come from identifying objects from the spatial relations in the scene. The same happens to scale context, once a contextual relationship establishes that objects have a limited set of size relations with other objects in the scene.

Therefore, the use of spatial and scale context involves using all forms of contextual information in the scene.

Over the years, many works have been proposed to extract a region of interest by using contextual information. But only after 2018 the first works that integrate the Biederman's contextual classes with CNNs were proposed. Leng et al. (2018a) have introduced the Context Learning Network (CLN), which aims to capture the pairwise relationships between objects and the global context of each one. By using the proposed method, the authors achieved 82.10%, 80.70%, and 38.40% mAPs on Pascal VOC 2007, Pascal VOC 2012, and Coco databases respectively. The CLN is composed of two subnetworks, a Multi-Layer Perceptron (MLP) which captures the pairwise relationships and a CNN that learn the global context of the image.

Leng et al. (2018b) also have proposed a context-aware U-Net, which aims to capture valuable contexts and improves the segmentation performance on ISBI Challenge Database. To do this a lightweight context transfer module was developed, to learn the rich context features present in the image, and the results obtained are similar to those presented by the state-of-the-art on the ISBI Challenge. To evaluate the results the authors have proposed three metrics: Warping Error, Rand Error, and Pixel Error, their respective values are 0.0001212%, 0.0212%, and 0.034600%.

Zhang et al. (2020) proposed one of the first works that made the use of spatial contextual information to object detection by using the Contextual Bidirectional Enhancement (CBD-E). CBD-E integrates the features of different background regions sequentially in two directions. On the one hand, a gate function is used to filter out unexpected information in the background and thus improve the recall of detection. On the other hand, a spatial-group-based visual attention mechanism is adopted to enhance the features of objects to reduce the false alarm. The gate function provides an approach to selecting meaningful information in the background, while the spatial-group-based visual attention mechanism enhances the information control ability of the gate function. To evaluate the proposed aproache the author employed the NWPU (Cheng et al., 2014), the RSOD (Long et al., 2017) and the DIOR (Li et al., 2020) datasets, achieving a mean Average Precision (mAP) score of 94.98%, 94.23% and 67.80% respectivelly.

Cai et al. (2021) have proposed a scale context based approach that combines the Cross-Channel Feature Pyramid Network (CFPN) with Foreground Attentive Detection Heads (FDH). In the presented method the scale information from different convolutional layers are extracted and incorporated in the network by using a cross-channel learning process. The proposed method achieves mAP score of 96.98%, 85.20% and 73.50% in the NWPU, RSOD and DIOR datasets respectively.

Liu et al. (2022) combined the scale and semantic contexts by joining the Balanced Feature Pyramid Network (BFPN) with a Task-interactive Head (TIH). In the proposed approach the BFPN balance the semantic and spatial information between high and shallow levels adoptive, and the TIH reduces the task misalignment between classification and regression. The proposed methods was evaluated in the HRRSD (Zhang et al., 2019) and DIOR datasets achieving a mAP score of 85.20% and 69.98% respectively.

3.3 FINAL REMARKS

In almost all works found in literature the authors presented the accuracy as the evaluation metric of the results of iris detection. However, in none of these except by the work proposed by Severo et al. (2018), were reported how the accuracy value is obtained. Because most of these studies did not report in detail how the metrics for evaluating the results were obtained, it is plausible to question the effectiveness of a proposed method about another.

Among the works available on the literature about Ocular Region Components (ORC) segmentation, was observed which most of these the segmentation result is evaluated by using the accuracy, which is obtained considering the total number of images in where the iris was correctly segmented in relation to the total images. While in the work proposed by Bezerra et al. (2018) the accuracy is calculated individually for each image then the general result is obtained by the mean value of all previously obtained accuracies. The individual values are obtained taking into account the correctly segmented pixels in relation to the total of it present on the pairwise segmentation mask. Bezerra et al. (2018) not only presented a pixel level based accuracy metric to evaluate the segmentation on iris scenario but also used the precision, recall and the Intersection over Union (IoU) to evaluate the segmented images.

After evaluating the metrics employed to evaluate the segmentation results of , we observed that the one that best quantifies the results obtained is IoU since by using it is possible to observe how close a segmented RoI, is near to ground truth segmented images.

Since we carried out an extensive verification of the works of extracting components from the ocular region, aiming to explore the best approaches developed and, consequently, the best metrics used to evaluate each of these approaches, we found that none of these explore the contextual relationships present in an image. However, the works that made use of this approach in the scene understanding scenario present promising results. In this way, it is opportune to develop a specific approach for the extraction of the RoI in the ocular region taking into account the contextual information. In the next chapter we presents the most relevant databases employed in the Ocular Region Components (ORC)-based biometrics.

4 DATABASES

In this chapter are presented the most relevant iris, and periocular region databases available on the literature. Section 4.1 present databases composed by images obtained in the Near-Infrared (NIR) wavelength. Section 4.2 present databases created by using Visible (VIS) wavelength sensor, cross-sensor, and cross-spectral images. Section 4.3 present databases composed by more than one type of data input such as iris, voice, frontal face images, and fingerprints. Table 3.2 summarizes all the databases described in this chapter.

4.1 NIR WAVELENGTH DATABASES

Databases composed by images captured at NIR wavelength are generally used to study the features present in the iris (CASIA, 2010; Phillips et al., 2008, 2010). However, also being possible to use them in another researches such as methodologies for the creation of synthetic irises (Shah and Ross, 2006; Zuo et al., 2007), vulnerabilities in iris recognition and liveness detection (Ruiz-Albacete et al., 2008; Czajka, 2013; Gupta et al., 2014; Kohli et al., 2016), impact of contact lenses on iris based biometrics (Baker et al., 2010; Kohli et al., 2013; Doyle et al., 2013; Doyle and Bowyer, 2015b), template aging in iris biometrics(Fenker and Bowyer, 2012; Baker et al., 2013), influence of alcohol consumption on iris recognition (Arora et al., 2012) and study of gender recognition through the iris (Tapia et al., 2016).

All the NIR databases presented in this section are proposed by Central Asia Student International Academic (CASIA). The first database proposed by them was the Central Asia Student International Academic - Iris V1 (CASIA-IrisV1). It has 756 images of 108 eyes with a resolution of 320×280 pixels, captured in two sections using a homemade iris camera (CASIA, 2002).

The CASIA-IrisV2 was the second database proposed by CASIA, and is composed by two subsets captured from two different sensors, OKI IRISPASS-h and CASIA-IrisCamV2. Each subset contains 1200 images from 60 classes with a resolution of 640×480 pixels (CASIA, 2004).

After the two previously version of databases made public available by the Central Asia Student International Academic, two more versions are developed, CASIA-IrisV3 and CASIA-IrisV4. CASIA-IrisV3 has a total of 22034 images from more than 700 subjects, arranged between three subsets: CASIA-Iris-Interval, CASIA-Iris-Lamp and CASIA-Iris-Twins. CASIA-IrisV4 is composed of six subsets, the three previously informed and three of the version itself. The three new subsets of CASIA-IrisV4 are CASIA-Iris-Distance, CASIA-Iris-Thousand and CASIA-Iris-Syn, which together with CASIA-IrisV3 subsets has 54601 iris, being those from more than 1800 genuine subjects and 1000 of virtual subjects. Each subset will be further detailed below, according to the specifications described by CASIA (2010).

Table 4.1: Datasets

Database	Summary	Controlled Environment	Iris Wavelength	Cross-sensor	Database Size subjects/images	Modality
CASIA-IrisV1 (CASIA, 2002)	Study of iris features captured in NIR	Yes	NIR	No	108 eyes / 756 images	Iris
CASIA-IrisV2 (CASIA, 2004)	Study of iris features captured in NIR	Yes	NIR	Yes	120 classes / 2400 images	Iris
UBIRIS v1 (Proença and Alexandre, 2005)	Iris segmentation and recognition in an uncontrolled environment	No	Visible	No	241 subjects / 1877 images	Iris
BioSec (Fierrez et al., 2007)	Create a large multimodal database	No	NIR	No	200 subjects / 3200 images	Iris / fingerprint / face / voice
CASIA-IrisV3-Interval (CASIA, 2010)	Study of the detailed texture features of iris images	Yes	NIR	No	249 subjects / 2639 images	Iris
CASIA-IrisV3-Lamp (CASIA, 2010)	Study problems of non-linear iris normalization and robust iris feature representation	Yes	NIR	No	411 subjects / 16212 images	Iris
CASIA-IrisV3-Twins (CASIA, 2010)	Study the dissimilarity and similarity between iris images of twins	Yes	NIR	No	200 subjects / 3183 images	Iris
CASIA-IrisV4-Distance (CASIA, 2010)	Can be used for multimodal biometrics (face, iris, periocular)	Yes	NIR	No	142 subjects / 2567 images	Iris
CASIA-IrisV4-Thousand (CASIA, 2010)	Study the uniqueness of iris features	Yes	NIR	No	1000 subjects / 20000 images	Iris
CASIA-IrisV4-Syn (CASIA, 2010)	Study the intra-class variation	N/A	N/A	N/A	1000 classes / 10000 images	Iris
UBIRIS v2 (Proenca et al., 2010)	Iris segmentation and recognition in an uncontrolled environment	No	Visible	No	261 subjects / 11102 images	Iris
MobBIOfake (Sequeira et al., 2014c)	Liveness detection in iris images obtained with mobile devices in uncontrolled environment	No	Visible	No	1600 images	Iris
MobBIO (Sequeira et al., 2014b)	Multimodal biometrics with mobile biometric systems	No	Visible	No	105 subjects / 1680 images	Iris / face / voice
PolyU (Nalla and Kumar, 2017a)	Study of iris recognition with cross-spectral images	N/A	Visible and NIR	Yes	209 subjects / 12540 images	Iris
MICHE-I (Marsico et al., 2015)	Recognition of iris images obtained by mobile devices in visible wavelength	No	Visible	Yes (mobile)	92 subjects / 3732 images	Iris
CSIP (Santos et al., 2015)	Biometric recognition in mobile environments using iris and periocular information	No	Visible	Yes (mobile)	50 subjects / 2004 images	Iris / periocular
VISOB (Rattani et al., 2016)	Periocular biometric recognition in visible spectrum with images captured by mobile devices	No	Visible	Yes (mobile)	550 subjects / 158136 images	Periocular
CROSS-EYED (Sequeira et al., 2016)	Recognition of iris and periocular region in cross-spectral images	No	Visible and NIR	Yes	120 subjects / 11520 images	Iris / periocular

The iris images of CASIA-Iris-Interval were captured with their own developed camera. The main characteristic of this database is that a circular near-infrared led illumination was used when the images were captured. In this way, this database can be used for studies on the detailing of texture features in iris images. This database is composed of 2639 images from 249 subjects and 395 classes, with a resolution of 320×280 pixels, obtained in two sections.

The CASIA-Iris-Lamp was collected using an unfixed sensor (OKI IRISPASS-h), thus, the individual collected the iris image with the sensor in their own hands. During the capture of the images, a lamp was switched on and off in order to produce more intra-class variations due to contraction and expansion of the pupil, creating a non-linear deformation. Therefore, this database can be used to study problems such as iris normalization and robust iris feature representation. A total of 16212 images with a resolution of 640×480 pixels, from 411 subjects and 819 classes, were collected in a single section.

During an annual twin festival in Beijing, Iris images from 100 pairs of twins were collected to create the CASIA-Iris-Twins database. It is an interesting database for the study of dissimilarity and similarity between iris images of twins. The database contains 400 classes of 200 subjects with 3183 images with a resolution of 640×480 pixels, captured using the OKI IRISPASS-h camera in a single section.

The CASIA-Iris-Distance database is composed of iris images captured using a long-range multi-modal biometric image acquisition and recognition system, developed by the authors of the database. The system can recognize users from a distance of up to 3 meters using a multi-camera intelligent imaging system with an active search for iris, face or palmprint patterns. The images contained in the database were captured with a high resolution camera, so a single image includes regions of interest for the feature extraction of both eyes and face. Detailed facial features such as skin pattern can also be used as a source of information for a multi-modal fusion. The database has 2567 images from 142 individuals and 284 classes with a resolution of 2352×1728 pixels.

The CASIA-Iris-Thousand contains 20000 iris images from 1000 subjects, with a resolution of 640×480 pixels, collected in a single section with the camera IKEMB-100, produced by IrisKing (IRISKING, 2017). Due to the number of individuals, this database can be used to study the uniqueness of iris features. The main sources of intra-class variations occurring due to specular reflections and eyeglasses.

The last subset of CASIA-IrisV4, called CASIA-IRIS-Syn, consists of a database with 10000 synthetic iris images from 1000 classes. The iris textures of these images were automatically synthesized from a CASIA-IrisV1 subset, using the segmentation approach described in (Tan et al., 2010). Intra-class variations were introduced in the images, such as deformation, blurring and rotation, making it difficult to represent iris features and matching.

4.2 VISIBLE WAVELENGTH, CROSS-SPECTRAL, AND CROSS-SENSORS IRIS IMAGES DATABASES

The iris databases composed by images captured on the VIS wavelength were created taking account that the set of problem that make use of images obtained using NIR wavelength under controlled environments can be considered solved (Proença and Alexandre, 2012).

Taking into account the results achieved by using images captured in NIR wavelength, investigations on biometric recognition with iris images obtained in uncontrolled environments and at visible wavelength began to be carried out (Proença and Alexandre, 2005; Proenca et al., 2010). There is also researches on biometric recognition using cross-spectral databases, i.e.,

with images of eyes of the same individual obtained at NIR and VIS wavelength (Hosseini et al., 2010; Sharma et al., 2014; Nalla and Kumar, 2017b; Algashaam et al., 2017).

Databases captured in controlled environments have few or none noise factors in the images. However, these conditions for capturing images are not easy to achieve and requires a high degree of collaboration from subjects. Within this context, the UBIRIS v1 database was created with the objective of providing images with different types of noise, simulating the capture of images with minimal collaboration from the users. The UBIRIS v1 database consists of 1877 images from 241 subjects, obtained in two sections, using the Nikon E5700 camera. For the first section (enrollment), noise factors were minimized, especially factors related to reflection, lighting and contrast. In the second section, the location for the capture was changed to introduce natural lighting factors and thus insert noise problems such as reflection, contrast, lighting and focus. The purpose of the second section was simulating an image capture with minimal or without active collaboration from the subjects. The database is available in three formats: color with a resolution of 800 x 600 pixels, color with 200×150 pixels and 200×150 pixel in grayscale (Proença and Alexandre, 2005).

The UBIRIS v2 database was built to represent the most realistic noise factors. For this reason, the images that constitute the database were obtained in visible spectrum without restrictions such as distance, angles, light and movement. The main purpose for the construction of this database was providing a tool for the research on the use of images at visible wavelengths, for the recognition of irises in an environment with adverse conditions. This database contains images captured with the Canon EOS 5D camera and resolution of 400×300 pixels in RGB, from 261 subjects containing 522 irises and a total of 11102 images (Proenca et al., 2010).

The MobBIOfake database was created with the aim to studying the liveliness detection of iris images obtained from mobile devices in uncontrolled environment (Sequeira et al., 2014c). It consists of 1600 fake iris images, obtained from a subset of 800 images belonging to the MobBIO database (Sequeira et al., 2014b). For the construction of the fake images, the original images were grouped by each subject and a pre-processing was done to improve the contrast. The images were printed using a professional printer in a high quality photo paper and recaptured using the same device and environment conditions used in the construction of MobBIO. Finally, the images were cropped and resized to fix dimensions.

The PolyU Cross-Spectral Iris database was developed to study iris recognition in cross-spectral images. The images from this database were obtained simultaneously under visible and NIR illumination. There are a total of 12540 iris images with a resolution of 640×480 pixels, with 15 images of each eye (left and right) in each spectrum from 209 subjects(Nalla and Kumar, 2017a).

In order to evaluate the state of the art in iris recognition with images from mobile devices, the Mobile Iris Challenge Evaluation (MICHE) competition (Part I) was created (Marsico et al., 2015). The MICHE-I or MICHEDB database consists of 3732 images from 92 subjects obtained by mobile devices in visible light. To simulate a real application, the iris images were obtained from the users themselves, indoors and outdoors, with and without glasses. Images of only one eye of each individual were captured. The mobile devices used and the resolution of the images obtained are the following: iPhone5 (1536×2048), Galaxy S4 (2322×4128) and Galaxy Tablet II (640×480). Due to the acquisition mode and the goal of creating the database, several noises are found in images such as specular reflections, focus, motion blur, occlusion due to eyelids, lighting variations, among others. There is a subset called MICHE FAKE, which has 80 printed iris images. These images were captured with iPhone5 and the Samsung Galaxy S4, then the images were printed with a LaserJet printer and captured again with the Samsung Galaxy S4. There is also another subset (MICHE Video) containing videos of irises from 10 people, obtained

in two modalities: indoor and outdoor. It were taken with two devices: Samsung Galaxy S4 and Samsung Galaxy Tab 2. In total, the subset has 120 videos of approximately 15 seconds each.

In order to gather images obtained in cross-sensor setup, simulating conditions found in mobile application scenarios, the CSIP (Cross-sensor iris and periocular dataset) database was created (Santos et al., 2015). To obtain the images, four different device models were used: Xperia Arc S (Sony Ericsson), iPhone 4 (Apple), w200 (Thl) and U8510 (Huawei). The resolutions of the images from each device are as follows: Xperia Arc S (Rear 3264×2448), iPhone 4 (Front 640×480, Rear 2592×1936), W200 (Front 2592×1936, Rear 3264×2448) and U8510 (Front 640×480, Rear 2048×1536). Combining the models with front and rear cameras and flash, 10 different setups were created with which the images were obtained. To simulate the variation of noise, the images were captured in different sites, with artificial, natural and mixed lighting conditions. The noise factors presented in the images are: multiple scales, chromatic distortions, image rotation, poor lighting, off-angle acquisition, out-of-focus images, deviated gaze and iris obstructions (including reflections) (Santos et al., 2015). The complete database has 2004 images of 50 subjects. The binary iris segmentation masks are also available, which were obtained using the method described by (Tan et al., 2010), winners of the NICE I competition.

The VISOB database was created for the ICIP 2016 Competition on mobile ocular biometric recognition, whose main objective was to evaluate the progress of research in the area of mobile ocular biometrics in the visible spectrum (Rattani et al., 2016). The front cameras of 3 mobile devices were used to obtain the images, such as: iPhone 5S at 720p resolution, Samsung Note 4 at 1080p resolution and Oppo N1 at 1080p resolution. The images that compose the database were captured in 2 sessions for each of the 2 visits, which occurred between 2 and 4 weeks, counting in the total 158136 images from 550 subjects. At each visit, it was required that each volunteer (subject) capture their own face using each of the three mobile devices at a distance between 8 and 12 inches from the face. For each session, images were captured under 3 light conditions: regular office light, offices lights off but dim ambient lighting still present (dim light) and next to sunlit windows (natural daylight settings). The collected database was preprocessed using the Viola-Jones eye detector and the region of the image containing the eyes was crop to a size of 240 x 160 pixels.

In order to investigate iris and periocular region recognition in cross-spectral images, the Cross-Spectral Iris/Periocular (Cross-Eyed) database was created (Sequeira et al., 2016). This database is composed of visible and NIR spectrum images obtained simultaneously with 2K x 2K pixel resolution cameras. There are 3 subsets: iris, periocular region without iris and ocular images obtained manually crop face images. Eight images of each eye were captured in each spectrum from 120 subjects, totaling 3840 images per subset. The periocular/ocular images have resolution of 900×800 pixels and the iris images, 400×300 pixels. All images were obtained at a distance of 1.5 meters, in an uncontrolled indoor environment, with a wide variation of ethnicity and eye colors, and lightning reflexes.

4.3 MULTIMODAL DATABASES

In addition to the databases database developed using iris/periocular region, there are some multimodal databases.

The BioSec baseline database consists of fingerprint images, frontal faces images, iris images and voice utterances. The data that compose the database was acquired from 200 subjects in two acquisition sessions. Environmental conditions such as lighting and background noise were not controlled, in order to simulate a real situation. The fingerprint images were obtained from three different sensors, the face images from a webcam and the voices were acquired with a

headset and a distant webcam microphone. For the iris data, 3200 images were captured, these being 4 images of each eye by acquisition, using the camera LG IrisAccess EOU3000 (Fierrez et al., 2007).

The MobBIO database was created due to the raising interest in mobile biometric applications and with the increasing interest and application of multimodal biometrics (Sequeira et al., 2014b). This database has biometric face, iris and voice data belonging to 105 subjects. The data were obtained with the mobile device Asus Transformer Pad TF 300T and for the captured of the images was used the rear camera of the device, version TF300T-000128 with 8 MP of resolution and autofocus. The iris images were obtained in two different lighting conditions, with varying eye orientations and occlusion levels. For each subject, 16 images (8 of each eye) were captured, which were cropped from a single image containing both eyes. Cropped images have a resolution of 300 x 200 pixels. Manual annotations of the iris and pupil contours are provided along with the database, but the iris illumination noises are not identified.

4.4 FINAL REMARKS

In this chapter are presented the most relevant databases used in works that explores problems related to periocular region components. In the next chapter are presented the approaches developed in this work to extract the Ocular Region Components (ORC) that is the central point of this this work.

5 PROPOSED METHODS AND OBTAINED RESULTS

In this chapter we describe the proposed methods taking into account the road-map presented in the Figure 1.6. Two lines of research were initially explored to cover the road-map stipulated for this work. During the initial period of this research, most of the efforts were concentrated on detecting the components of the periocular region. All the proposed approaches and results for detection can be seen in Section 5.1. Once the ocular region components detection approaches were explored and the results obtained in this line of research surpassed state-of-the-art, we redirected our efforts to the ocular region components segmentation, and all the approaches presented to this scenario and the obtained results are presented in Section 5.2. Finally, in Section 5.3.1 are presented the final remarks about the proposed approaches present in this Chapter.

5.1 OCULAR REGION COMPONENTS DETECTION

In this section, the approaches for detecting the ORC developed in this work are presented. The research explored the detection of two complementary regions,, the iris and the ocular region as a whole. In the first moment, we explore iris detection, comparing traditional object detection methods with strategies based on convolutional neural networks (see Section 5.1.1). After validating the hypothesis that it was or was not possible to detect the iris with convolutional neural networks, we investigated the simultaneous detection of the iris and the periocular region, the proposed approaches and the obtained results can be seen in Section 5.1.2. Finally, in Section 5.1.3 the final remarks about the detection approaches are presented.

5.1.1 Iris Detection

This approach defines the iris location problem as the delimitation of the smallest rectangular window that encompasses the iris region. In this scenario, we proposed two window based detectors. Section 5.1.1.1 present a sliding window detector based on features from Histogram of Oriented Gradients (HOG) and a linear Support Vector Machines (SVM) classifier, and Section 5.1.1.2 present a deep learning based detector fine-tuned from YOLOv2 object detector. The results obtained by using these approaches can be seen in Section 5.1.1.3.

5.1.1.1 Histogram of Oriented Gradients and Support Vector Machines

Despite image acquisition with different devices, lighting conditions, variations of translation, rotation and scale (Zhu et al., 2000), the iris presents a standard structure, following patterns of texture, shape and edge orientations, which can be described by a feature descriptor and interpreted by a classifier.

Histogram of Oriented Gradients (HOG) is a feature descriptor used in computer vision for object detection. This method quantizes the gradient orientation occurrences in regions of an image, extracting shape information from objects (Dalal and Triggs, 2005). Figure 5.1 illustrates an image described by HOG.

In this approach, each window was divided into blocks of 2×2 cells, and cells of 8×8 pixels. For each cell, the horizontal and vertical gradients in all pixels are calculated. Thus, the orientations and magnitudes of the gradient are obtained. The gradient orientations are then quantified in nine directions (i.e., bins).

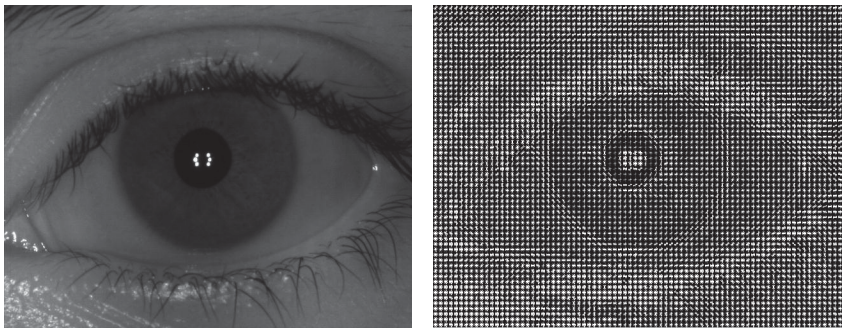
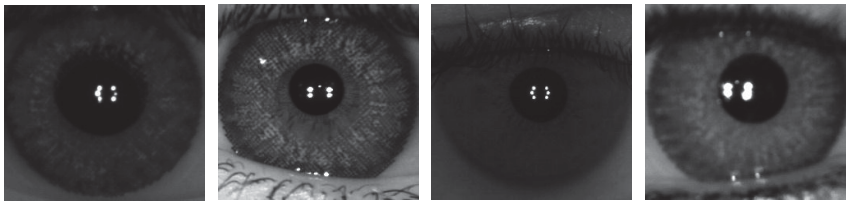
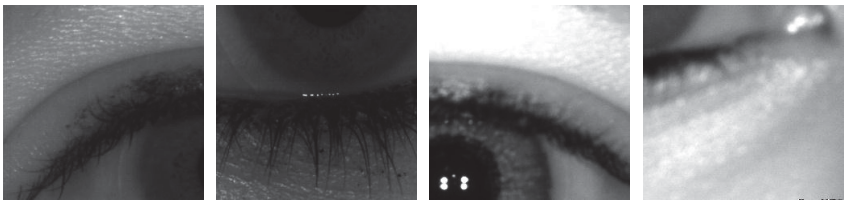


Figure 5.1: Exemple of image described by HOG.

In order to avoid effects of light and contrast variation, the histograms of all cells on blocks (2×2 cells) are normalized. Observe that those block have an overlap of 50%. The HOG feature vector that describes each iris window is then constructed by concatenating the normalized cell histograms for all blocks. Finally, a feature vector (2×2 blocks $\times$ 8 cells $\times$ 9 orientations) is obtained to describe each iris candidate window.



(a) Positive samples



(b) Negative samples

Figure 5.2: Training samples used by SVM.

The window containing the iris region (ground truth) from each training image is extracted and used to compose the examples of positive windows. Furthermore, windows that are completely outside or have only a small intersection with the iris region are extracted and considered negative windows. We created 10 negative windows for each positive window, which are used for training a SVM classifier using a linear kernel. Figures 5.2(a) and 5.2(b) illustrate, respectively, positive and negative samples used for the training of the proposed approach.

The SVM was first presented by Boser et al. (1992), and it is one of the most used classification methods in recent years (Franchi et al., 2016; Ruiz et al., 2014). To find the decision boundary, the SVM minimizes the upper limit of the generalization error, which is obtained by maximizing the margin distance from the training data.

In order to perform the iris location, a sliding window approach with different scales is applied to each test image. We adopted windows with size 50×50 pixels as canonical scale. From this scale, we used 6 lower scales and 8 higher scales by a factor of 5%. The image region that presents the greatest similarity with the iris can be found using the decision border generated by the SVM, which will return the highest positive response for the best estimated iris location.

5.1.1.2 You Only Look Once (YOLO) Object Detector

Currently, deep Convolutional Neural Networks (CNNs) are one of the most efficient ways to perform image classification, segmentation and object detection. In this approach, we evaluate the concepts proposed by Redmon and Farhadi (2017a) in the Darknet, which is an open source neural network framework used to implement YOLOv2, a state-of-the-art real-time object detection system (Redmon et al., 2016a).

The YOLOv2 network, as most CNNs, is composed of three main operation layers to object detection, which are: convolution, max pooling and classification, the latter occurs using fully connected layers. On Darknet, convolutional layers work as feature extraction, in other words, a convolutional kernel is sliding in the input image. The network architecture is inspired by the GoogLeNet model for image classification (Szegedy et al., 2015). The original YOLO has 24 convolutional layers that produce different feature maps from the input. The feature maps are then processed by max pooling layers, which dimensionally reduces the previously obtained feature map. That is, max pooling divides the feature map into blocks and reduces each block into one value. Instead of the inception modules used by GoogLeNet, YOLO uses 1×1 reduction layers followed by 3×3 convolutional layers, similar to Lin et al. (2013).

In this approach we use a fast version of YOLO (i.e., Fast-YOLO), a similar neural network with fewer convolutional layers (15 instead of 24) and fewer filters in those layers. Other than the size of the network, all training and testing parameters are the same for both YOLO and Fast-YOLO. The Proposed architecture is shown in Table 5.1.

Table 5.1: Fast-YOLO network used to detect the periocular region. We reduced the number of filters in the last convolutional layer from 125 to 30 in order to output 1 class instead of 20.

	Layer	Filters	Size	Input	Output
0	conv	16	$3 \times 3/1$	$416 \times 416 \times 1/3$	$416 \times 416 \times 16$
1	max		$2 \times 2/2$	$416 \times 416 \times 16$	$208 \times 208 \times 16$
2	conv	32	$3 \times 3/1$	$208 \times 208 \times 16$	$208 \times 208 \times 32$
3	max		$2 \times 2/2$	$208 \times 208 \times 32$	$104 \times 104 \times 32$
4	conv	64	$3 \times 3/1$	$104 \times 104 \times 32$	$104 \times 104 \times 64$
5	max		$2 \times 2/2$	$104 \times 104 \times 64$	$52 \times 52 \times 64$
6	conv	128	$3 \times 3/1$	$52 \times 52 \times 64$	$52 \times 52 \times 128$
7	max		$2 \times 2/2$	$52 \times 52 \times 128$	$26 \times 26 \times 128$
8	conv	256	$3 \times 3/1$	$26 \times 26 \times 128$	$26 \times 26 \times 256$
9	max		$2 \times 2/2$	$26 \times 26 \times 256$	$13 \times 13 \times 256$
10	conv	512	$3 \times 3/1$	$13 \times 13 \times 256$	$13 \times 13 \times 512$
11	max		$2 \times 2/1$	$13 \times 13 \times 512$	$13 \times 13 \times 512$
12	conv	1024	$3 \times 3/1$	$13 \times 13 \times 512$	$13 \times 13 \times 1024$
13	conv	1024	$3 \times 3/1$	$13 \times 13 \times 1024$	$13 \times 13 \times 1024$
14	conv	30	$1 \times 1/1$	$13 \times 13 \times 1024$	$13 \times 13 \times 30$
15	detection				

5.1.1.3 Experimental Protocol and Results

The experiments presented in this section compares the classical and outstanding Daugman's iris location approach with two window based detectors: 1) a sliding window detector based on features from HOG and a linear SVM classifier; 2) a deep learning based detector fine-tuned

from YOLO V2 object detector. Four scenarios (intra-sensor, inter-sensor, multiple-sensors and mixing of databases) were proposed to evaluate the results obtained by using the previously described approaches.

Six databases were employed on the previously described experiments: IIIT-D CLI (Kohli et al., 2013), NDCLD15 (Doyle and Bowyer, 2015a), MobBIOfake (MobBIOfake) (Sequeira et al., 2014a), NDCCL (Doyle and Bowyer, 2014), Central Asia Student International Academic - Iris -Interval (CASIA-Interval) (CASIA, 2010) and BERC mobile-iris database (Kim et al., 2016). Except the by the NDCLD15, all other databases were manually annotated. The NDCLD15 annotations were provided by the database authors (Doyle and Bowyer, 2015a) (more details about the employed image databases can be seen in Chapter 4).

5.1.1.3.1 Evaluation protocol

In order to analyze the experiments, we employ the following metrics: Recall, Precision, Accuracy and IoU. These metrics are defined between the area of the ground truth and predicted bounding boxes in terms of False Positives (FP), False Negatives (FN), True Positives (TP), and True Negatives (TN) pixels, and can be formally expressed bellow.

$$Recall = \frac{TP}{TP + FN} \tag{5.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.3}$$

$$IoU = \frac{TP}{FP + TP + FN} \tag{5.4}$$

5.1.1.3.2 Intra-sensor Results

Table 5.2 shows the results obtained by intra-sensor experiments, in other words, experiments in which the models were trained and tested with images from the same sensor. The YOLO V2 Convolutional Neural Network (CNN) based framework achieved the best averages in almost all analyzed metrics and required less processing time for iris location per image. The exception is for CASIA IrisV3 Interval database where Daugman's method presented slightly better Precision (96.23% against 96.02%) and Accuracy (97.38% against 97.10%). This surprising result can be explained by the high level of cooperation and control in the image acquisition of such database. That is, the Daugman's method takes somehow advantage of the scenario. Anyway, the YOLO CNN locates the iris in real-time (0.02 seconds per image, on average) using our fast Titan XP GPU, whilst the Daugman's method and the HOG-SVM approach demand, on average, 3.5 and 5.2 seconds, respectively, to locate the iris in each image using a single CPU core.

5.1.1.3.3 Inter-sensor Results

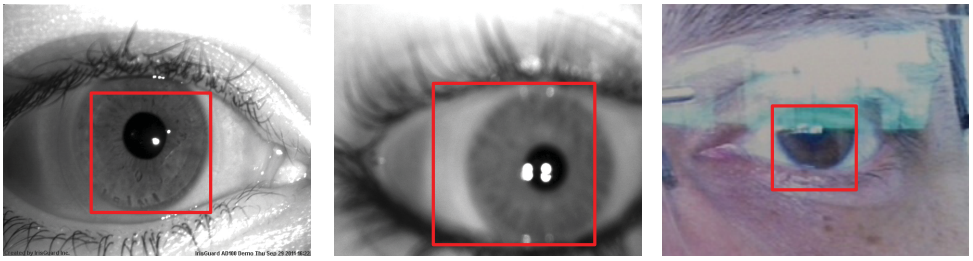
Table 5.2: Iris detection results by using the intra-sensor approach (%)

Database		Recall			Precision			Accuracy			IoU		
		Daugman	HOG	YOLO	Daugman	HOG	YOLO	Daugman	HOG	YOLO	Daugman	HOG	YOLO
NDCCL	AD100	84.60	92.39	**98.78**	82.49	94.78	**95.03**	94.28	96.98	**98.49**	80.41	87.52	**93.84**
	LG4000	93.41	96.72	**97.81**	92.15	90.80	**97.73**	97.53	97.24	**99.05**	89.67	87.76	**95.06**
IIIT-D CLI	Vista	85.49	94.51	**97.85**	89.34	92.24	**93.71**	95.38	98.10	**98.28**	80.82	87.23	**91.76**
	Cogent	86.24	**96.44**	96.02	92.82	87.99	**95.58**	96.34	96.67	**98.33**	82.61	84.76	**91.84**
MobBIO	Real	76.32	95.77	**96.81**	74.71	72.26	**94.02**	85.26	95.33	**98.97**	70.79	68.76	**91.02**
	Fake	75.81	93.28	**96.06**	73.45	74.33	**95.05**	84.81	95.26	**98.90**	70.12	68.99	**91.27**
BERC		88.19	92.83	**98.10**	85.64	87.95	**93.56**	98.72	98.49	**99.71**	79.10	85.10	**91.15**
CASIA-Interval		96.38	96.97	**97.79**	**96.23**	88.48	96.02	**97.38**	92.21	97.10	90.95	86.17	**91.24**
NDCLD15		91.63	96.04	**97.28**	89.76	90.29	**95.71**	96.67	97.14	**98.54**	85.34	86.85	**93.25**

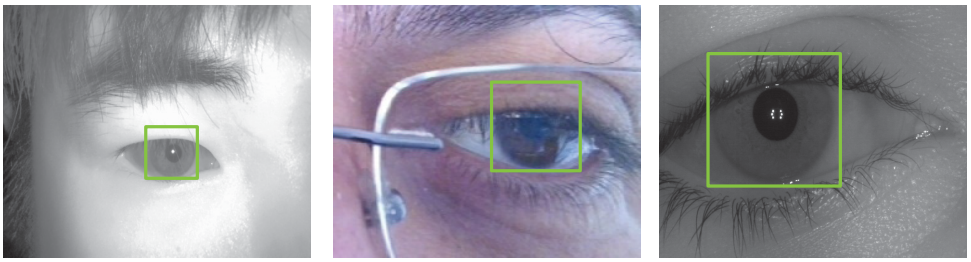
In addition, for databases containing images acquired with more than one sensor, inter-sensor experiments were performed and are presented in Table 5.3. That is, we train the detectors with images of one sensor and test/evaluate then on the images from other sensor. These experiments show that in some cases YOLO CNN did not achieve promising results as previously shown. For example, in the database NDCCL, when fine tunning/training the detector with images from the AD100 sensor and testing with the ones from LG4000 sensor.

Table 5.3: Inter-sensor results (%)

Database	Set		Recall		Precision		Accuracy		IoU	
	Train	Test	HOG-SVM	YOLO	HOG-SVM	YOLO	HOG-SVM	YOLO	HOG-SVM	YOLO
NDCCL	AD100	LG4000	**92.95**	79.25	**91.13**	89.18	**96.84**	92.67	**85.78**	68.71
	LG4000	AD100	93.22	**97.99**	93.15	**93.59**	96.78	**97.94**	86.76	**91.63**
IIIT-D CLI	Vista	Cogent	**96.89**	96.13	89.89	**94.21**	96.43	**97.98**	83.94	**90.57**
	Cogent	Vista	93.44	**98.26**	**93.61**	87.97	**97.08**	96.65	**87.55**	80.92



(a)



(b)

Figure 5.3: Samples of iris location obtained in the experiments: (a) poor results due to a homogeneous training set; (b) good results achieved with images of different sensors on training set.

The reason for the poor result might lie in the fact that the database for that specific sensor (AD100) has only 600 images, thus not allowing a good generalization of the trained CNN. In Figure 5.3(a), we can observe some examples where the iris location obtained by the YOLO method did not achieved good results.

5.1.1.3.4 Multiple-sensors Results

In order to better analyze and understand the results of the inter-sensor experiments and to confirm our hypothesis that the YOLO's poor performance is due to few/homogeneous training samples, experiments were performed combining images from multiple sensors of the same databases. The results obtained in this new experiment can be seen in Table 5.4. It highlights the importance of a diverse collection of images for the training set in CNNs. With a larger number of images acquired from different sensors in the training set, the CNN was able to better generalize, increasing the correct iris location in most cases. Some examples of good iris location can be seen in Figure 5.3(b).

Table 5.4: Combined sensor results (%), same databases

Database	Set		Recall		Precision		Accuracy		IoU	
	Train	Test	HOG-SVM	YOLO	HOG-SVM	YOLO	HOG-SVM	YOLO	HOG-SVM	YOLO
NDCCL	AD100 & LG4000	LG4000	95.37	**99.29**	92.93	**99.68**	97.48	**99.77**	88.63	**98.91**
	AD100 & LG4000	AD100	91.77	**99.37**	94.77	**97.42**	96.85	**99.36**	86.91	**96.85**
IIIT-D CLI	Vista & Cogent	Cogent	96.73	**97.26**	87.15	**96.48**	96.50	**98.49**	84.17	**92.50**
	Vista & Cogent	Vista	94.20	**98.34**	92.74	**93.79**	97.01	**98.55**	87.41	**91.78**

5.1.1.3.5 Mixing Databases results

Table 5.5 contains the results obtained by experiments where YOLO was trained with the training sets of all the databases and tested in the test images of all the databases. The results achieved by the Daugman's method applied to all the test images are also presented, and we used specific parameters for each database. By analyzing these figures, we observe that YOLO strikingly outperforms the Daugman's method in all analyzed metrics.

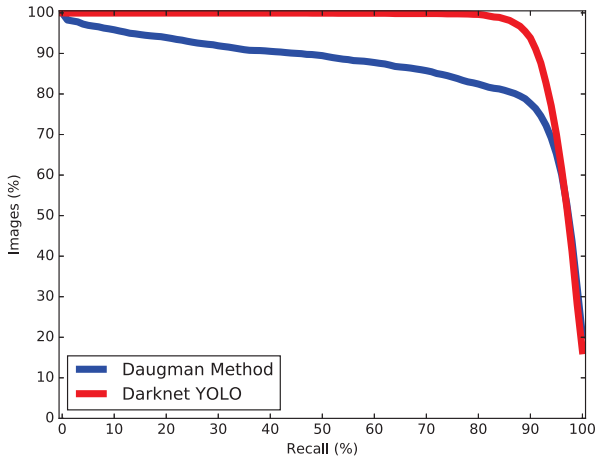


Figure 5.4: Recall curve of both Daugman and YOLO methods applied to all test sets.

Figure 5.4 shows the behavior of the recall curve for the experiment reported in Table 5.5. It depicts how the percentage of images varies when we required a minimum Recall rate. These

curve highlights how YOLO is a promising alternative to iris location, since all tested images achieved Recall values above 80%. That is, at least 80% of the required region of a iris is certainly located by the YOLO detector.

Table 5.5: Combined sensor results (%), mixed databases

Method	**Set**		**Recall**	**Precision**	**Accuracy**	**IoU**	**Time**
	Train	Test					
YOLO	All training sets	All test sets	**97.13**	**95.20**	**98.32**	**92.54**	**0.02 s**
Daugman	-	All test sets	86.45	86.28	94.04	81.09	3.50 s

5.1.2 Simultaneous Iris and Ocular Region Detection

Taking into account the fact of the iris appears as one of the main biological characteristics in security systems since it remains unchanged over time and its uniqueness level is high (Zhu et al., 2000). Furthermore, the identification using the iris region is non-invasive, that is, there is no need for physical contact to obtain and analyze an iris image (Jain et al., 1998). However, after decades of research in personal identification, it has been observed that better results can be achieved by combining different biometric modalities (Nigam et al., 2015; Tan and Kumar, 2012; Chang et al., 2004). A good example of it is the combination of iris and periocular-based biometrics (Xiao et al., 2012; Marsico et al., 2017).

In this method, we compare the detection of the iris and periocular regions being performed separately and simultaneously using two well-known object detection networks: YOLOv2 (Redmon and Farhadi, 2017b) and Faster R-CNN (Ren et al., 2015). Such deep models were chosen due to the fact that (i) promising results were recently obtained using them in other detection tasks (Ding et al., 2018; Ko and Sim, 2017; Laroca et al., 2018); and (ii) handcrafted features are easily affected by noise and might not be robust for unconstrained scenarios.

Typically, in biometric systems that use iris and/or periocular region images as input, the first step in which efforts should be applied is the detection of the RoI (Daugman, 1993), as a poor detection would probably impair the effectiveness of the subsequent steps of the system (Daugman, 2004; Tan and Kumar, 2012). Recently, Zanlorensi et al. (Zanlorensi et al., 2018) showed that impressive iris recognition rates can be achieved when using deep representations having as system input the bounding boxes of the iris region, without the iris segmentation preprocessing. Also using deep representations and having as input a squared region (i.e., a bounding box), Luz et al. (Luz et al., 2018) achieved state-of-the-art results for periocular recognition. Such results, shorter execution times compared to single detection approaches (in which the iris and the periocular region are detected separately), and the promising results obtained in preliminary experiments support our motivation to detect both regions simultaneously.

Currently, one of the most accurate ways to perform image classification, segmentation and object detection is using deep CNNs. Therefore, we propose the simultaneous detection of the iris and periocular regions using two object detection models: YOLOv2 (Redmon and Farhadi, 2017b) (Section 5.1.2.1) and Faster Region-based Convolutional Network (Faster R-CNN) + Feature Pyramid Network (FPN) (Ren et al., 2015) (Section 5.1.2.2). It should be noted that (i) we trained both models from scratch; (ii) such models were chosen because promising results were obtained using them in other detection tasks (Ding et al., 2018; Ko and Sim, 2017; Laroca et al., 2018). The results obtained by using these approaches can be seen in Section 5.1.2.3.

5.1.2.1 You Only Look Once (YOLO)

Table 5.6 presents the YOLOv2 model, employed for detecting the iris and the periocular region. The architecture has 19 convolutional and 5 max-pooling layers. The convolutional layers, except for the last one, are divided into two groups: external and internal. The layers belonging to the external group use kernels of size 3×3, whereas the layers belonging to the internal group use kernels of size 1×1. Alternating 1×1 convolutional layers reduce the features space from preceding layers (Redmon et al., 2016b). The convolutional blocks are composed of: convolution, batch normalization, and a Rectified Linear Unit (ReLU).

Table 5.6: The YOLOv2 model, modified for the detection of the iris and the periocular region. There are 30 filters in the last convolutional layer when the regions are detected separately and 35 when they are detected simultaneously.

#	Layer	Group	Filters	Size	Input	Output	#	Layer	Group	Filters	Size	Input	Output
0	conv	External	32	$3 \times 3/1$	$416 \times 416 \times 1/3$	$416 \times 416 \times 32$	12	conv	External	512	$3 \times 3/1$	$26 \times 26 \times 256$	$26 \times 26 \times 512$
1	max			$2 \times 2/2$	$416 \times 416 \times 32$	$208 \times 208 \times 32$	13	conv	Internal	256	$1 \times 1/1$	$26 \times 26 \times 512$	$26 \times 26 \times 256$
2	conv	External	64	$3 \times 3/1$	$208 \times 208 \times 32$	$208 \times 208 \times 64$	14	conv	External	512	$3 \times 3/1$	$26 \times 26 \times 256$	$26 \times 26 \times 512$
3	max			$2 \times 2/2$	$208 \times 208 \times 64$	$104 \times 104 \times 64$	15	conv	Internal	256	$1 \times 1/1$	$26 \times 26 \times 512$	$26 \times 26 \times 256$
4	conv	External	128	$3 \times 3/1$	$104 \times 104 \times 64$	$104 \times 104 \times 128$	16	conv	External	512	$3 \times 3/1$	$26 \times 26 \times 512$	$26 \times 26 \times 512$
5	conv	Internal	64	$1 \times 1/1$	$104 \times 104 \times 128$	$104 \times 104 \times 64$	17	max			$2 \times 2/2$	$26 \times 26 \times 512$	$13 \times 13 \times 512$
6	conv	External	128	$3 \times 3/1$	$104 \times 104 \times 64$	$104 \times 104 \times 128$	18	conv	External	1024	$3 \times 3/1$	$13 \times 13 \times 512$	$13 \times 13 \times 1024$
7	max			$2 \times 2/2$	$104 \times 104 \times 128$	$52 \times 52 \times 128$	19	conv	Internal	512	$1 \times 1/1$	$13 \times 13 \times 1024$	$13 \times 13 \times 512$
8	conv	External	256	$3 \times 3/1$	$52 \times 52 \times 128$	$52 \times 52 \times 256$	20	conv	External	1024	$3 \times 3/1$	$13 \times 13 \times 512$	$13 \times 13 \times 1024$
9	conv	Internal	128	$1 \times 1/1$	$52 \times 52 \times 256$	$52 \times 52 \times 128$	21	conv	Internal	512	$1 \times 1/1$	$13 \times 13 \times 1024$	$13 \times 13 \times 512$
10	conv	External	256	$3 \times 3/1$	$52 \times 52 \times 128$	$52 \times 52 \times 256$	22	conv	External	1024	$3 \times 3/1$	$13 \times 13 \times 512$	$13 \times 13 \times 1024$
11	max			$2 \times 2/2$	$52 \times 52 \times 256$	$26 \times 26 \times 256$	23	conv		30/35	$1 \times 1/1$	$13 \times 13 \times 1024$	$13 \times 13 \times 30/35$

As this model does not have fully connected layers, it can receive images of any size as input. We adopted an input size of 416×416 pixels due to the good results achieved employing these dimensions in (Redmon and Farhadi, 2017b). We also reduced the number of filters in the last convolutional layer to match our number of classes. The number of filters in that layer is given by

$$filters = (C + 5) \times A \,, \tag{5.5}$$

where A is the number of anchor boxes (we use $A = 5$) used to predict bounding boxes and C is the number of classes, in our case either $C = 1$ or $C = 2$ to detect the iris and periocular regions separately or simultaneously, respectively. Thus, there are 30 filters in the last convolutional layer when the regions are detected separately and 35 when they are detected simultaneously.

The main difference between the YOLOv2 model proposed in (Redmon and Farhadi, 2017b) and the one used in this work is that we removed the route layers, i.e., layers that concatenate a list of previous layers together. In preliminary experiments, we observed that removing such layers did not negatively affect the results obtained in our tasks and also reduced the execution time.

5.1.2.2 Faster R-CNN + Feature Pyramid Network

We employ the Faster R-CNN model (Ren et al., 2015) combined with a FPN (Lin et al., 2017) to detect the Ocular Region Components (ORC). The architecture of the proposed approach can be seen in Figure 5.5. Faster R-CNN is commonly composed of (i) a feature map extraction network; (ii) a region proposal network and (iii) a detection network. We replaced the standard CNN feature extraction module by an FPN, and thus multiple feature map layers are generated with better quality information than the regular implementation of Faster R-CNN.

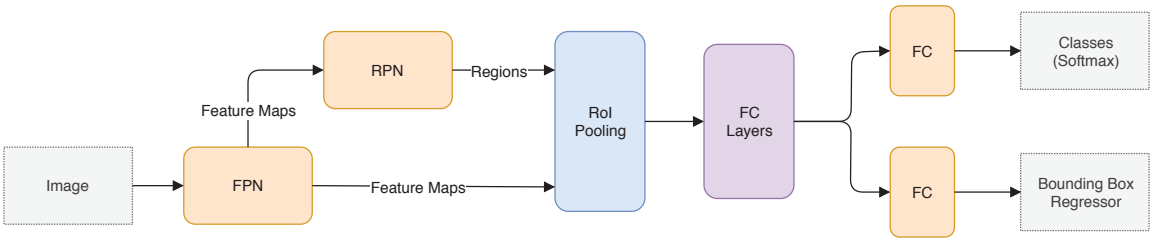


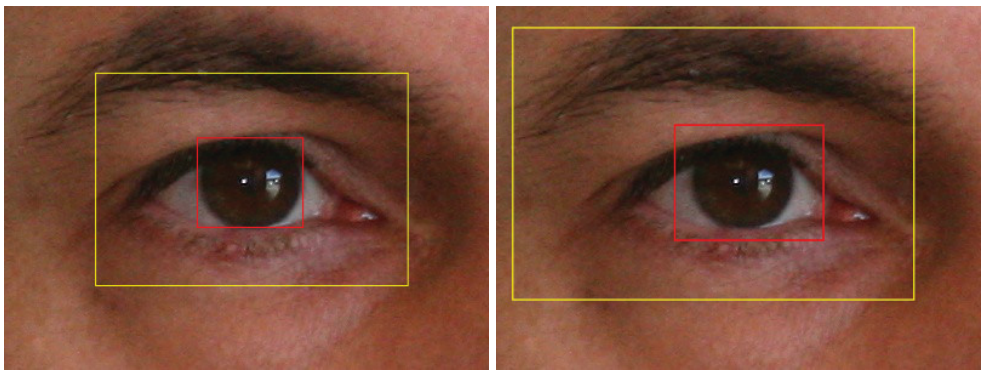
Figure 5.5: Faster R-CNN + FPN architecture overview.

5.1.2.3 Experimental Protocol and Results

In this section are presented coarse annotation concept (Section 5.1.2.3.1), the experimental protocol (Section 5.1.2.3.2) and the obtained results (Section 5.1.2.3.3).

5.1.2.3.1 Coarse Annotations

To perform the experiments described in this, we propose the use of coarse annotations both to train and to evaluate our networks. As can be seen in Fig. 5.6, we define as a coarse annotation the region around the RoI so that the edges of the bounding box remain outside the limits of the fine annotations proposed by Severo et al. (Severo et al., 2018). More specifically, the delimited region is larger than the one typically used in fine annotations, and the iris is not well-centered. Also, in some cases, the eyebrows were left out the RoI, as the images from some databases used in this work do not contain that region.



(a) Fine (b) Coarse

Figure 5.6: Examples of fine and coarse annotations of both the iris (red) and the periocular region (yellow).

It is worth noting that the coarse annotations were made manually by two volunteers and that no strict rules of how annotations should be made were defined (besides simple instructions and the fact that were coarse and not fine annotations). Hence, there are random variations (in size, position, aspect ratio, etc.) among annotations of different images.

We believe coarse annotations can be used in recognition systems based on the iris and/or the periocular region, given the much smaller engineering effort required to manually annotate the training images. In other words, we conjecture that deep models for person identification may achieve promising results even when these regions are not perfectly segmented.

We employed the following public databases: CASIA-Iris-Interval (CASIA, 2010), CASIA-Iris-Lamp (CASIA, 2010), CASIA-Iris-Thousand (CASIA, 2010), Cross-Eyed-VIS (Sequeira et al., 2014a), CSIP (Santos et al., 2015), MICHE-I (Marsico et al., 2015), MobBIO (Sequeira et al., 2014a), PolyU-VIS (Nalla and Kumar, 2017a), UBIRIS.v2 (Proenca et al., 2010)

and VISOB (Rattani et al., 2016) (More information about the employed databases can be seen in 4). These databases were chosen because they are widely used in the biometric recognition literature (Zanlorensi et al., 2018; Silva et al., 2018; Aginako et al., 2016; Deshpande et al., 2014), which we plan to investigate in future works.

5.1.2.3.2 Evaluation Protocol

The evaluation of an automatic detection approach is performed in a pixel-to-pixel comparison between the ground truth and the predicted bounding boxes. Therefore, we use the mean F-score, IoU and mAP evaluation metrics. Following Severo et al. (Severo et al., 2018), to first compute the precision and recall metrics and then the F-score, we consider as correct the bounding boxes detected with an IoU value above 0.5 with the ground truth. This bounding box evaluation, defined in the PASCAL VOC Challenge (Everingham et al., 2010), is interesting since it penalizes both over- and under-estimated objects.

It is worth noting that we use coarse annotations as the ground truth, as the databases do not provide fine annotations of the position of the iris and periocular regions on each image. In this sense, instead of evaluating the predicted bounding boxes in relation to the exact location of the iris/periocular region, we evaluated how close to the ground truth it is.

In order to perform a fair evaluation and comparison of the proposed approaches, we divided each database into three subsets, being 40% of the images for training, 40% for testing and 20% for validation. We adopt this protocol (i.e., with a larger test set) to provide more samples for analysis of statistical significance. Also, in the statistical direction, we perform the Wilcoxon signed-rank test (Wilcoxon et al., 1970) to verify if there is a statistical difference between the detection approaches.

5.1.2.3.3 Results and Discussion

In this section are presented the results obtained by using the experimental protocol presented in Section 5.1.2.3.2. To compare the proposed approaches, we report the F-score values in order to analyze the trade-off between precision and recall measures, however, we focus on the IoU metric since we want to assess how close are the predicted bounding boxes compared to the ground truth.

When analyzing the results regarding ***iris*** detection (see top of Table 5.7), in 10 of 11 experiments the highest mean IoU value was achieved using Faster R-CNN. In general, the best results were obtained when simultaneously detecting the iris and the periocular region. The exceptions are in the CASIA-Iris-Lamp, Cross-Eyed-VIS, MICHE-I, and UBIRIS.v2 databases, where detecting both regions separately performed better, probably due to the fact that there are not many variations in iris and periocular region arrangement in the images of these databases. However, as the difference in the results obtained with both approaches is very small, we applied the Wilcoxon signed-rank test and observed that there is no statistical difference between detecting the iris and the periocular region simultaneously or separately in the CASIA-Iris-Lamp, Cross-Eyed-VIS, CSIP and MobBIO databases. In this way, in Table 5.7, we highlighted (light gray) the results obtained in these databases.

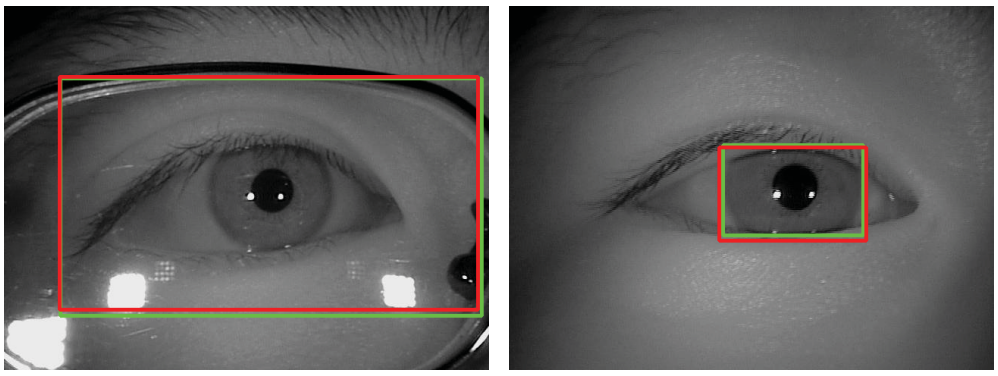
Similar behavior occurred in the detection of the ***periocular*** region, however, in this case, all the best results were attained employing the Faster R-CNN model. In this scenario, the detection results using the single-class detection approach CSIP, UBIRIS.v2 and VISOB

databases presented the best values. Similar to the results on iris detection, the difference between the IoU values attained between the approaches is close and there is no statistical difference in the CASIA-Iris-Thousand and Cross-Eyed-VIS databases and that result was also highlighted in Table 5.7.

Table 5.7: Detection results. In the table the Single and Multi columns present the results obtained when detecting the iris and periocular regions separately and simultaneously, respectively. The values in bold represent the highest IoU values obtained, while the highlighted results indicate the cases in which there is no statistical difference according to the Wilcoxon statistical tests.

Database	F-score				IoU (%)			
	YOLOv2		Faster R-CNN		YOLOv2		Faster R-CNN	
	Multi	Single	Multi	Single	Multi	Single	Multi	Single
Iris								
CASIA-Iris-Interval	0.90	0.92	0.97	0.96	82.81	86.20	**94.77**	93.98
CASIA-Iris-Lamp	0.96	0.96	0.98	0.95	92.38	93.06	96.08	**97.31**
CASIA-Iris-Thousand	0.97	0.98	0.98	0.98	95.71	94.39	**97.72**	97.58
Cross-Eyed-VIS	0.92	0.92	0.94	0.94	85.79	86.45	90.39	**90.44**
CSIP	0.92	0.73	0.95	0.95	87.97	58.12	**91.61**	91.55
MICHE-I	0.88	0.83	0.92	0.92	80.32	72.07	86.27	**86.48**
MobBIO	0.95	0.95	0.96	0.96	91.52	91.40	**94.14**	93.79
PolyU-VIS	0.91	0.86	0.94	0.94	**93.81**	76.32	89.12	89.31
UBIRIS.v2	0.89	0.89	0.91	0.91	81.16	81.75	85.16	**85.26**
VISOB	0.91	0.89	0.96	0.96	85.04	81.32	**93.09**	92.80
Periocular Region								
CASIA-Iris-Interval	0.96	0.98	0.98	0.98	92.65	96.19	**97.80**	96.79
CASIA-Iris-Lamp	0.98	0.97	0.99	0.98	97.15	96.02	**98.08**	97.71
CASIA-Iris-Thousand	0.97	0.98	0.99	0.99	95.92	96.44	**98.19**	98.19
Cross-Eyed-VIS	0.92	0.92	0.96	0.96	86.86	86.89	**92.74**	92.56
CSIP	0.95	0.95	0.87	0.96	91.61	91.76	84.97	**92.96**
MICHE-I	0.85	0.85	0.90	0.90	75.88	74.97	**83.66**	83.51
MobBIO	0.96	0.96	0.97	0.97	94.21	94.09	**95.50**	94.83
PolyU-VIS	0.96	0.87	0.98	0.98	93.57	77.95	**96.74**	96.41
UBIRIS.v2	0.87	0.88	0.91	0.91	78.98	80.03	85.19	**85.44**
VISOB	0.93	0.94	0.97	0.98	87.17	89.11	96.08	**96.35**

We emphasize that most of the best results were obtained using the Faster R-CNN + FPN approach, which we believe to be justified by the fact that FPNs perform a better feature map extraction compared to other approaches (Lin et al., 2017).



(a) Periocular Region (b) Iris Detection

Figure 5.7: Best iris and periocular region detection performed by the Faster R-CNN simultaneous detection approach. The green bounding boxes represent the coarse annotations, while the red ones represent the detected regions.

It should be noted that the IoU values obtained were higher than 95% for both iris and periocular region detection in the databases where the images were captured using a NIR sensor. These results were achieved by using the Faster R-CNN simultaneous detection approach, and the better detected iris and periocular region can be seen in Figure 5.7.

Despite the promising results, it is necessary to observe that the obtained IoU presented values lower than 90% in some scenarios. The lower IoU values were found in databases in which the images were captured using more than one sensor with no preprocessing (i.e., MICHE-I and CSIP) and in those composed with lower quality images (i.e., UBIRIS.v2). By analyzing these images, we can understand what made the results obtained by the approaches on these databases below than 90% of IoU: i) the use of eyeglasses; ii) the presence of more than one eye.

5.1.3 Final Remarks

The Ocular Region Components (ORC) location is a preliminary but extremely important task in specific applications such as iris recognition, spoofing and liveness detection, as well as contact lens detection, among others. In this section, we evaluate two ORC detection scenarios. In the first scenario, we evaluated if a Convolutional Neural Network (CNN) based detection approach can outperform a Histogram of Oriented Gradients (HOG) based system. And by using the CNN-based method, we achieved state-of-the-art results in iris detection.

Once we identified the potential of using neural networks to detect periocular region components, we decided to test the proposed approach in a more challenging scenario, the simultaneous detection of iris and periocular region using coarse annotations, as seen in Section 5.1.2. In the new ORC detection scenario both the presented CNN based approach have presented excellent results, however, the Faster R-CNN + FPN approach showed slightly better results. Even though the approach based on the YOLO architecture presents slightly lower results than the Faster R-CNN + FPN method in accuracy terms, the frame rate offered by this surpasses the first approach by 200%.

By analyzing the presented results in both detection scenarios, we have defined that the YOLO-based detection framework will be employed as the preprocessing approach in the Ocular Region Components (ORC) segmentation investigation presented in the next section. The performed experiments by using the detection methods presented in this section a

5.2 OCULAR REGION COMPONENTS SEGMENTATION

In this section, the approaches to segment the Ocular Region Components (ORC) developed in this work are presented. In the first moment, we explore proposed segmentation protocol (see Section 5.2.1)). After, we explore sclera segmentation scenario, comparing state-of-the-art segmentation CNN-based approaches (see Section 5.2.2). Once we verify the feasibility of segmenting the sclera using neural networks, we also investigate the iris segmentation problem, as can be seen in Section 5.2.3. Finally in Section 5.2.4 the final remarks about the segmentation approaches are presented.

5.2.1 Segmentation Protocol

In a Ocular Region Components (ORC) based recognition system, some images might present specular highlights in regions of the subject's face, which affect the conventional segmentation approaches employed in this scenario, as we can observe in our preliminary tests. Many of these regions were erroneously classified as sclera. In this way we propose a two stage segmentation protocol as can be seen in Figure 5.8. The proposed two-stage protocol extracts the Region of

Interest (RoI) by using a Convolutional Neural Network (CNN) based object detection approach, and the output RoI of it is used as input to the segmentation approaches. Details about the RoI extration approach can be seen in Section 5.2.1.1, and the segmentation approaches used in this protocol are presented in Section 5.2.1.2.

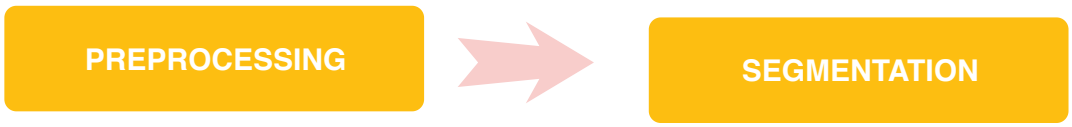


Figure 5.8: Segmentation approaches flow.

5.2.1.1 Preprocessing

We employ as Ocular Region Detection (ORD) approach the object detection method presented in Section 5.1.2.1. The ORD was trained using the iris coarse annotations concept described in Section 5.1.2.3.1. Once the computational model training is finished, we can extract the Ocular region by applying padding (2 times the iris radius to the width and 1 time to the height) in the detected patch (i.e., iris), as can be seen in Figure 5.9.

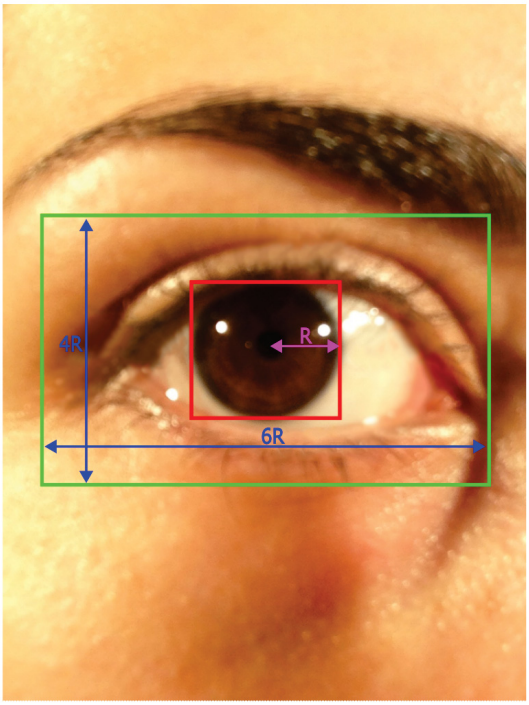
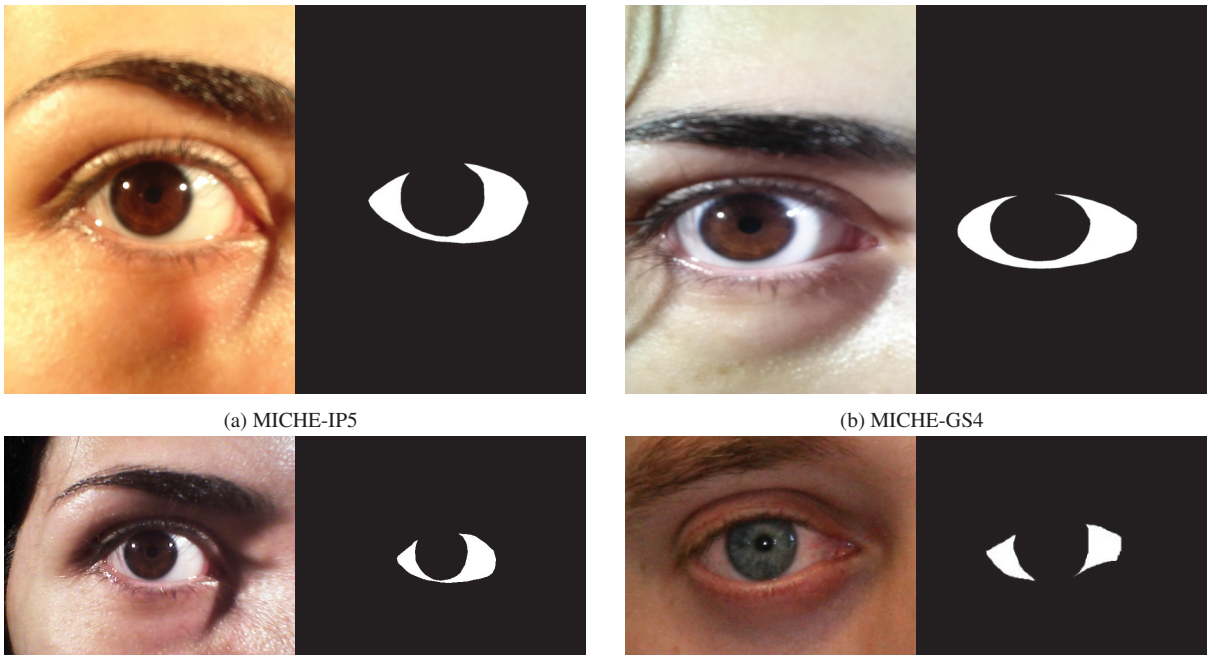


Figure 5.9: Exemple of image described by RoI.

After detecting the ocular region, only the RoI was maintained in each image. Figure 5.10 shows an example of each subset of the original image (without ORD) and the segmentation mask created by us.

Figure 5.11 shows, four cropped images after ORD and their respective masks. It is noteworthy that most specular highlights are removed after ORD. At last, the RoI is resized according to each segmentation approach presented in Section 5.2.1.2. The input sizes were chosen based on the original architectures of the approaches (see Table 5.8).



(a) MICHE-IP5 (b) MICHE-GS4

(c) MICHE-GT2 (d) UBIRIS.v2

Figure 5.10: Four examples of the masks created by us.

Table 5.8: Image dimensions used in each approach.

Approach	Image - Dimension	Mask - Dimension
FCN	$320 \times 240 \times 3$	$320 \times 240 \times 1$
GAN	$256 \times 256 \times 3$	$256 \times 256 \times 3$
SegNet	$320 \times 240 \times 3$	$320 \times 240 \times 1$

(a) MICHE-IP5 (b) MICHE-GS4

(c) MICHE-GT2 (d) UBIRIS.v2

Figure 5.11: Periocular regions detected and replicated to the masks.

5.2.1.2 Segmentation Approaches

In this thesis, we proposed two new approaches to sclera segmentation based on CNNs, one based on FCN (Teichmann et al., 2016b) (see Section 5.2.1.2.1) and another one based on GAN (Isola et al., 2016) (Section 5.2.1.3). To the best of our knowledge, both kinds of networks have never been studied in the Ocular Region Components (ORC) segmentation scenario. FCN is used for

segmentation in a large range of applications, from medical to satellite image analysis (Henry et al., 2018; Roth et al., 2018), while GAN was employed initially in scene understanding task (Luc et al., 2016).

5.2.1.2.1 Fully Convolutional Network

Fully Convolutional Network (FCN) are deep neural networks in which an image is provided as input and a mask is generated at the output. This mask is a binary image (of the same size) where each pixel is classified as iris or not iris. Basically, we employed the MultiNet (Teichmann et al., 2016a) segmentation decoder without the classification and detection decoders. The encoder consists of the first 13 layers of the VGG-16 network (Simonyan and Zisserman, 2014). The features extracted from its fifth pooling layer were then used by the segmentation decoder, which follows the FCN architecture (Long et al., 2015) (see Fig. 5.12).

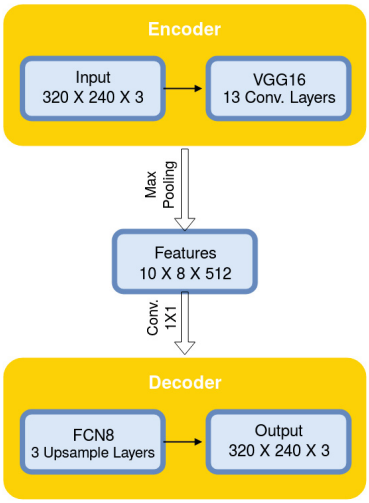


Figure 5.12: FCN architecture for sclera segmentation.

The fully-connected layers of the VGG-16 network were transformed into 1×1 convolutional layers to produce a low-resolution segmentation. Then, three transposed convolution layers were used to perform up-sampling. Finally, high-resolution features were extracted through skip layers from lower layers to improve the up-sampled results.

The pre-trained VGG-16 weights on ImageNet were used to initialize the encoder, the segmentation decoder, and the transposed convolutional layers. The training is based on the Adam optimizer algorithm (Kingma and Ba, 2014), with the following parameters: learning rate of 10^{-5}, dropout probability of 0.5, weight decay of 5^{-4} and standard deviation of 10^{-4} to initialize the skip layers.

5.2.1.3 Generative Adversarial Network (GAN):

Generative Adversarial Networks (GANs) are deep neural networks composed by both generator and discriminator networks, pitting one against the other. First, the generator network receives noise as input and generates samples. Then the discriminator network receives samples of training data and those of the generator network, being able to distinguish between the two sources (Goodfellow et al., 2014). The GAN architecture for iris segmentation is shown in Fig. 5.13.

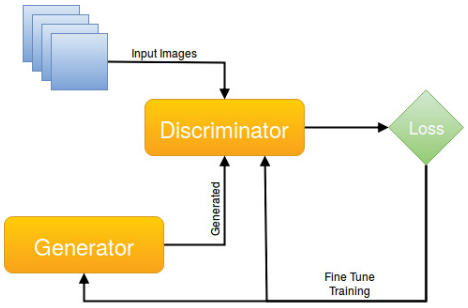


Figure 5.13: GAN architecture for sclera segmentation.

Basically, the generator network learns to produce more realistic samples throughout each iteration, while the discriminator network learns to better distinguish the real and synthetic data.

Isola et al. (Isola et al., 2016) presented the GAN approach used in this work, which is a Conditional Generative Adversarial Network (CGAN) able to learn the relation between an image and its label, and from that, generate a variety of image types, which can be employed in various tasks such as photo-generation and semantic segmentation.

5.2.2 Sclera Segmentation

As described in previous sections the segmentation is one of the first steps in which efforts should be applied in a reliable sclera-based recognition system. Incorrect segmentation can either reduce the region of blood vessels or introduce new patterns such as eyelashes and eyelids, impairing the system effectiveness (Das et al., 2013; Delna et al., 2016). Trying to reduce the incorrect segmented region on sclera scenario in this section we evaluate the use of the segmentation approaches presented in Section 5.2.1.2, by using the protocol presented in Section 5.2.2.1.

5.2.2.1 Experimental protocol and Results

In this section are presented baseline employed in sclera segmentation scenario (see Section 5.2.2.1.1), the datasets used in the experiments (see Section 5.2.2.1.2), the evaluation protocol (see Section 5.2.2.1.3) and the obtained results (see Section 5.2.2.1.4).

5.2.2.1.1 Baseline

We choose the Encoder-Decoder (ED) as baseline of sclera segmentation taking into account the fact of this approach achieved state-of-the-art results in *Sclera segmentation and eye recognition benchmarking competition* (Das et al., 2017). The Encoder-Decoder (ED), also called autoencoder, is a neural network trained in order to copy its input to the output. The purpose is to learn data encoding which can be used for dimensionality reduction or even for file compression (Audebert et al., 2017).

The ED (SegNet) used in this work was presented in (Badrinarayanan et al., 2017). SegNet consists of a stack of encoders followed by a corresponding stack of decoders which feed a soft-max classification layer. Decoders map low-resolution features extracted by encoders to an image with the same dimension as the input. The architecture used is presented in Table 5.9.

Table 5.9: SegNet architecture.

Layer		Filters	Size	Input	Output	Layer		Filters	Size	Input	Output
1	enc	64	3×3	$320 \times 240 \times 3/1$	$320 \times 240 \times 64$	19	up		2×2	$10 \times 8 \times 512$	$20 \times 15 \times 512$
2	enc	64	3×3	$320 \times 240 \times 64$	$320 \times 240 \times 64$	20	dec	512	3×3	$20 \times 15 \times 512$	$20 \times 15 \times 512$
3	max		2×2	$320 \times 240 \times 64$	$160 \times 120 \times 64$	21	dec	512	3×3	$20 \times 15 \times 512$	$20 \times 15 \times 512$
4	enc	128	3×3	$160 \times 120 \times 64$	$160 \times 120 \times 128$	22	dec	512	3×3	$20 \times 15 \times 512$	$20 \times 15 \times 512$
5	enc	128	3×3	$160 \times 120 \times 128$	$160 \times 120 \times 128$	23	up		2×2	$20 \times 15 \times 512$	$40 \times 30 \times 512$
6	max		2×2	$160 \times 120 \times 128$	$80 \times 60 \times 128$	24	dec	512	3×3	$40 \times 30 \times 512$	$40 \times 30 \times 512$
7	enc	256	3×3	$80 \times 60 \times 128$	$80 \times 60 \times 256$	25	dec	512	3×3	$40 \times 30 \times 512$	$40 \times 30 \times 512$
8	enc	256	3×3	$80 \times 60 \times 256$	$80 \times 60 \times 256$	26	dec	256	3×3	$40 \times 30 \times 512$	$40 \times 30 \times 256$
9	enc	256	3×3	$80 \times 60 \times 256$	$80 \times 60 \times 256$	27	up		2×2	$40 \times 30 \times 256$	$80 \times 60 \times 256$
10	max		2×2	$80 \times 60 \times 256$	$40 \times 30 \times 256$	28	dec	256	3×3	$80 \times 60 \times 256$	$80 \times 60 \times 256$
11	enc	512	3×3	$40 \times 30 \times 256$	$40 \times 30 \times 512$	29	dec	256	3×3	$80 \times 60 \times 256$	$80 \times 60 \times 256$
12	enc	512	3×3	$40 \times 30 \times 512$	$40 \times 30 \times 512$	30	dec	128	3×3	$80 \times 60 \times 256$	$80 \times 60 \times 128$
13	enc	512	3×3	$40 \times 30 \times 512$	$40 \times 30 \times 512$	31	up		2×2	$80 \times 60 \times 512$	$160 \times 120 \times 128$
14	max		2×2	$40 \times 30 \times 512$	$20 \times 15 \times 512$	32	dec	128	3×3	$160 \times 120 \times 128$	$160 \times 120 \times 128$
15	enc	512	3×3	$20 \times 15 \times 512$	$20 \times 15 \times 512$	33	dec	64	3×3	$160 \times 120 \times 128$	$160 \times 120 \times 64$
16	enc	512	3×3	$20 \times 15 \times 512$	$20 \times 15 \times 512$	34	up		2×2	$160 \times 120 \times 64$	$320 \times 240 \times 64$
17	enc	512	3×3	$20 \times 15 \times 512$	$20 \times 15 \times 512$	35	dec	64	3×3	$320 \times 240 \times 64$	$320 \times 240 \times 64$
18	max		2×2	$20 \times 15 \times 512$	$10 \times 8 \times 512$	36	dec	2	3×3	$320 \times 240 \times 64$	$320 \times 240 \times 2$

5.2.2.1.2 Databases

The experiments were carried out in two subsets of well-known iris databases: UBIRIS.v2 and MICHE-I. An overview of both subsets can be seen in Table 5.10. Remark that we do not use the SSBC (Das et al., 2015) and SSRBC (Das et al., 2016) databases in our experiments as only the test sets were made available by the authors.

Table 5.10: Overview of the databases used in this work. All of these are a subset of the original database.

Database	Images	Subjects	Resolution
UBIRIS.v2	500	261	400×300
MICHE-I	1,000	92	Various
MICHE-GS4	333	92	Various
MICHE-IP5	323	92	Various
MICHE-GT2	344	92	640×480

5.2.2.1.3 Evaluation protocol

The performance evaluation of an automatic segmented mask is performed in a pixel-to-pixel comparison between the ground truth and the predicted image. Therefore, we use the following metrics: Precision, Recall and F-score.

To perform a fair evaluation and comparison of the proposed approaches in all databases, we divided each into three subsets, being 40% of the images for training, 40% for testing and 20% for validation.

5.2.2.1.4 Results and Discussions

In this section we evaluate the results obtained by using the sclera segmentation approaches presented in Section 5.2.2. The results obtained by both the baseline (SegNet) and the

proposed approaches are shown in Table 5.11. The baseline presented considerably worse results. We believe this is due to the size of the training set, since SegNet was originally employed in a large dataset (Radu et al., 2015). Radu et al. (Radu et al., 2015) generated a dataset with 54,000 images using data augmentation. However, we did not have access to the database used by them, and thus a more direct comparison with their methodology was not possible to be done.

Table 5.11: Results achieved using the baseline and the proposed protocols.

Database	Approach	Recall %	Precision %	F-score %
UBIRIS.v2	GAN	87.48 ± 08.19	87.10 ± 08.16	86.82 ± 05.88
	SegNet	72.48 ± 17.15	87.52 ± 08.53	77.82 ± 13.08
	FCN	**87.31 ± 06.68**	**88.45 ± 06.98**	**87.48 ± 03.90**
MICHE-I	GAN	87.07 ± 10.81	86.39 ± 12.02	86.27 ± 09.97
	SegNet	64.59 ± 24.73	83.39 ± 18.53	69.87 ± 22.34
	FCN	**87.59 ± 11.28**	**89.90 ± 09.82**	**88.32 ± 09.80**
MICHE-GS4	GAN	85.72 ± 12.53	86.12 ± 13.01	85.20 ± 11.31
	SegNet	66.50 ± 26.34	76.09 ± 23.80	67.92 ± 23.87
	FCN	**88.24 ± 12.03**	**88.65 ± 10.62**	**88.12 ± 10.56**
MICHE-IP5	GAN	88.11 ± 07.40	87.71 ± 07.71	87.42 ± 05.43
	SegNet	31.90 ± 26.05	79.40 ± 32.93	40.95 ± 29.19
	FCN	**87.51 ± 11.61**	**89.32 ± 05.22**	**87.80 ± 08.24**
MICHE-GT2	GAN	86.20 ± 15.02	83.81 ± 15.73	84.50 ± 14.28
	SegNet	73.77 ± 21.20	76.46 ± 18.29	72.33 ± 18.26
	FCN	**87.86 ± 12.23**	**88.50 ± 12.68**	**87.94 ± 11.59**

Better results were obtained using the proposed approaches. In the UBIRIS.v2 subset, the GAN-based sclera segmentation attained a F-score value of 86.82% (±5.88), while the approach based on FCN achieved 87.48% (±3.90). Although there is little difference between the F-score values obtained by both methods, the standard deviation presented when using FCN was slightly lower than when GAN was employed for the segmentation.

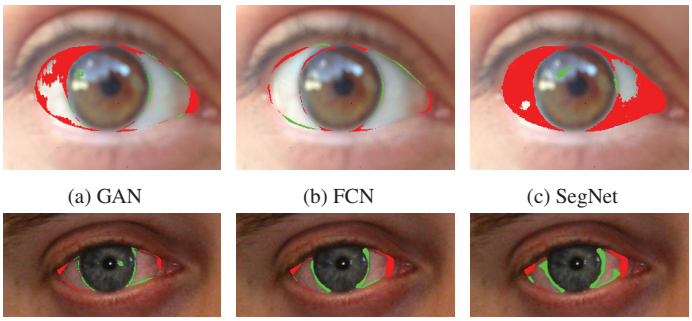
The same happened in all subsets used in our experiments, fact that makes us believe that the FCN approach is best suited for sclera segmentation. However, the results obtained with the GAN-based segmentation should not be diminished, since they were very close to the best results.

Here we perform a visual analysis. For this task, we randomly chose an image from the UBIRIS.v2 subset. Figure 5.14 demonstrate an outcome in the segmentation of the sclera. As can be seen, the FCN approach presented a considerably better segmentation result when compared to the baseline and GAN. It is noteworthy the consistency presented with FCN-based segmentation technique, observed in all sclera images generated in this work.

5.2.3 Iris Segmentation

The identification of individuals based on their biological and behavioral characteristics has a higher degree of reliability compared to other means of identification, such as passwords or access cards. Several characteristics of the human body can be used for person recognition (e.g., face, signature, fingerprints, iris, sclera, retina, voice, etc.) (Jain et al., 2016). The characteristics present in the iris make it one of the most representative and safe biometric modalities. This circular diaphragm forming the textured portion of the eye is capable of distinguishing individuals with a high degree of uniqueness (Wildes, 1997; Jain et al., 2004).

As described in (Jillela and Ross, 2015), an automated biometric system for iris recognition is composed of four main steps: (i) image acquisition, (ii) iris segmentation, (iii) normalization and (iv) feature extraction and matching. The segmentation consists of



(a) GAN (b) FCN (c) SegNet

(d) GAN (e) FCN (f) SegNet

Figure 5.14: Samples of segmented sclera using the ground truth for highlighting errors: green and red pixels represent the False Positive (FP) and False Negative (FN) respectively.

locating and isolating the iris from other regions (e.g., the sclera, surrounding skin regions, etc.), therefore it is the most critical and challenging step of the system. Incorrect segmentation usually affects the subsequent steps, impairing the system performance (Rattani and Derakhshani, 2017).

Leveraging the advent of CNNs we propose the use of the approaches presented in Section 5.2.1.2 for iris segmentation task. The experimental protocol employed to evaluate the proposed approaches can be observed in Section 5.2.3.1.

5.2.3.1 Experimental protocol and Results

In this section are presented baseline employed in sclera segmentation scenario (see Section 5.2.2.1.1), the datasets used in the experiments (see Section 5.2.2.1.2), the evaluation protocol (see Section 5.2.2.1.3) and the obtained results (see Section 5.2.2.1.4).

5.2.3.1.1 Datasets

The experiments were carried out on well-known and challenging publicly available iris datasets with both NIR and VIS images. Being these: BioSec (BioSec): (Fierrez et al., 2007); Central Asia Student International Academic - Iris - Interval v3 (CasiaI3) (Tan and Sun, 2005); Central Asia Student International Academic - Iris - Thousand v4 (CasiaT4) (Tan et al., 2010); Cross-Spectral Iris/Periocular (CrEye-Iris) (Sequeira et al., 2016); and MICHE-I (Marsico et al., 2015). More details about the employed datasets can be see in Chapter 4.

5.2.3.1.2 Baselines

We selected three baseline frameworks described (and available) in the literature to compare with our approaches with: Open Source Iris Recognition System Version 4.1 (OSIRISv4.1), Iris Segmentation Framework (IRISSEG) and Haindl & Krupička (Haindl and Krupička, 2015).

The **OSIRISv4.1** (Othman et al., 2016a) framework is composed of four key modules: segmentation, normalization, feature extraction and matching. Nevertheless, we used only the segmentation module to compare it with our method. Although the performance of this framework was only reported in datasets with NIR images, we applied it on both NIR and VIS

image datasets. This framework has input parameters such as minimum/maximum iris diameter. For a fair comparison, we tuned the parameters for each dataset in order to obtain the best results.

The **IRISSEG** (Gangwar et al., 2016a) framework was designed specifically for non-ideal irises and is based on adaptive filtering, following a coarse-to-fine strategy. The authors emphasize that this approach does not require adjustment of parameters for different datasets. As in OSIRISv4.1, we report the performance of this framework on both NIR and VIS images.

The **Haindl & Krupička** (Haindl and Krupička, 2015) framework was used to evaluate the results achieved by the proposed approach on VIS datasets. This method was developed for colored eyes images obtained through mobile devices and used as the baseline in the MICHE-II (Marsico et al., 2017) contest. We did not report the Haindl & Krupička (Haindl and Krupička, 2015) performance on NIR images datasets since it was not possible to generate the segmentation masks using the executable provided by the authors.

5.2.3.1.3 Evaluation Protocol

A pixel-to-pixel comparison between the ground truth (manually labeled) and the algorithm prediction (i.e., the mask/segmentation) generate an average segmentation error E computed as a pixel divergence, given by the exclusive-or logical operator $\otimes$ (i.e., XOR) (Proença and Alexandre, 2012), denoted by

$$E = \frac{1}{h \times w} \sum_{i} \sum_{j} M_k(i,j) \otimes GT_k(i,j)\,, \tag{5.6}$$

where i and j are the coordinates in the mask M and ground truth GT images, h and w stand for the height and width of the image, respectively. Lower and higher E values represent better and worse results, respectively. We also reported the F-Measure (F1) measure which is a harmonic average of Precision and Recall (Jalilian et al., 2017).

In order to perform a fair evaluation and comparison of the proposed methodologies to the baselines in all datasets, we randomly divided each dataset into two subsets, containing 80% of the images for training and the remainder for evaluation. The stopping learning criteria was 32,000 iterations.

5.2.3.1.4 Results and Discussions

The experiments were performed the protocol proposed in Section 5.2.2.1.3. Moreover, in order to analyze the robustness among sensors from the same environment (i.e., NIR or VIS) of the proposed FCN and GAN approaches, they were training using either all NIR or VIS image datasets and then evaluated on the same scenario. Finally, a visual and qualitative analysis showing some good and poor results is performed.

We trained and tested the FCN and GAN approaches on each dataset to compare them with the benchmarks. Table 5.12 shows the results obtained when using the proposed evaluation protocol (see Section 5.2.3.1.3).

Remark that both IRISSEG and OSIRISv4.1 frameworks presented good results in NIR datasets, always reaching F1 values over 90%. Nonetheless, our proposed approaches presented statistically better F1 values for all datasets even in the NIR datasets, which are the IRISSEG and OSIRISv4.1 specific image domain. Observe that there are no results for the approach by Haindl & Krupička (Haindl and Krupička, 2015) since it was not developed for NIR images.

Table 5.12: Iris segmentation results using the proposed protocol.

Dataset	Method	F1 %	E %
BioSec (NIR)	OSIRISv4.1 (Othman et al., 2016b)	92.62 ± 03.19	01.21 ± 00.47
	IRISSEG (Gangwar et al., 2016b)	93.94 ± 05.88	01.06 ± 01.20
	FCN Proposed	**97.46 ± 00.74**	**00.44 ± 00.12**
	GAN Proposed	**96.82 ± 02.83**	**00.74 ± 01.40**
	FCN Proposed	**97.90 ± 00.68**	**01.15 ± 00.37**
	GAN Proposed	**96.13 ± 05.35**	**01.45 ± 03.71**
CasiaT4 (NIR)	OSIRISv4.1 (Othman et al., 2016b)	87.76 ± 08.01	01.34 ± 00.64
	IRISSEG (Gangwar et al., 2016b)	91.39 ± 08.13	00.95 ± 00.54
	FCN Proposed	**94.42 ± 07.54**	**00.61 ± 00.58**
	GAN Proposed	**95.38 ± 03.72**	**01.40 ± 00.93**
Cross-Spectral Iris/Periocular (CrEye-Iris) (VIS)	OSIRISv4.1 (Othman et al., 2016b)	46.53 ± 29.25	13.22 ± 06.33
	IRISSEG (Gangwar et al., 2016b)	61.72 ± 33.55	10.58 ± 10.38
	Haindl & Krupička (Haindl and Krupička, 2015)	76.81 ± 23.73	05.69 ± 04.58
	FCN Proposed	**97.04 ± 01.21**	**00.96 ± 00.36**
	GAN Proposed	**92.61 ± 05.86**	**03.02 ± 03.22**
MICHE-I (VIS)	OSIRISv4.1 (Othman et al., 2016b)	33.85 ± 35.86	01.99 ± 02.90
	IRISSEG (Gangwar et al., 2016b)	19.34 ± 33.03	01.90 ± 03.37
	Haindl & Krupička (Haindl and Krupička, 2015)	63.12 ± 33.30	01.32 ± 02.10
	FCN Proposed	**83.01 ± 19.47**	**00.37 ± 00.43**
	GAN Proposed	**87.42 ± 13.08**	**03.27 ± 03.13**

Looking at VIS datasets, the results obtained were slightly worse than in the NIR datasets. This is because VIS images usually have more noise, e.g., reflections. The best F1 and E values achieved for the VIS datasets were achieved by the FCN approach with 97.04%(±01.21) and 00.37%(±00.43), respectively, in the CrEye-Iris and MICHE-I datasets.

It is worth noting that the FCN approach is the one with the smallest E values in almost all scenarios. This result can be explained by the fact that the FCN approach took advantage of transfer learning, while the GAN approach was trained from scratch.

5.2.3.1.5 Suitability and Robustness

Here, experiments for evaluating the suitability and robustness of the proposed approaches are presented. By suitability, we expect that models trained with a specific kind of images, i.e. NIR or VIS images, work as well as when training on a specific dataset. By robustness, we expect that models trained with all kind of images (NIR and VIS) perform as well as when training on a specific dataset.

In summary, the suitability is evaluated by training the models using only NIR or VIS images (i.e., FCN and GAN trained on the NIR merged and VIS also merged datasets). The robustness is evaluated by training the models using all images available (NIR and VIS merged). The results are presented in Tables 5.13 and 5.14, respectively. Note that we report the results of the separate test subsets as well, to facilitate visual comparison between the tables.

By comparing the values presented in Table 5.13 with those reported in Table 5.12, we can observe that the values vary slightly, and thus we can state that the proposed approaches are stable in the suitability scenario.

When comparing the results presented in Table 5.13 and Table 5.14, we noticed that the obtained values of $F1$ and E were similar in NIR datasets. On the other hand, the performance was considerably lower in VIS datasets. Therefore, the proposed approaches are robust for both NIR and VIS images. However, the GAN approach presented a decrease in the results, while the FCN obtained little variation.

Table 5.13: Suitability (bold lines) for NIR and VIS environments.

Dataset	Method	F1 %	*E* %
BioSec	FCN	97.24 ± 00.81	00.58 ± 00.30
	GAN	90.19 ± 05.52	02.22 ± 01.39
CasiaI3	FCN	97.43 ± 00.74	00.55 ± 00.29
	GAN	97.10 ± 01.83	00.75 ± 01.10
CasiaT4	FCN	95.87 ± 02.66	01.25 ± 00.67
	GAN	82.65 ± 13.98	05.52 ± 04.15
NIR	**FCN**	**96.69 ± 01.43**	**00.78 ± 00.63**
	GAN	**94.04 ± 07.93**	**01.72 ± 02.69**
CrEye-Iris	FCN	96.71 ± 01.11	01.12 ± 00.80
	GAN	93.21 ± 02.30	01.88 ± 00.53
MICHE-I	FCN	88.36 ± 11.88	01.90 ± 02.20
	GAN	89.49 ± 06.76	03.11 ± 02.24
VIS	**FCN**	**89.56 ± 12.36**	**02.40 ± 02.21**
	GAN	**92.58 ± 04.89**	**02.80 ± 02.05**

Table 5.14: Robustness (bold lines) of the iris segmentation approaches.

Dataset	Method	F1 %	*E* %
BioSec	FCN	96.57 ± 01.14	00.70 ± 00.24
	GAN	85.48 ± 07.63	03.45 ± 01.97
CasiaI3	FCN	97.69 ± 00.82	00.50 ± 00.33
	GAN	93.33 ± 01.98	00.87 ± 00.92
CasiaT4	FCN	95.39 ± 03.20	01.46 ± 01.12
	GAN	85.68 ± 12.92	03.98 ± 02.80
NIR	FCN	96.89 ± 06, 60	00.82 ± 00.59
	GAN	89.87 ± 07.93	02.39 ± 01.78
CrEye-Iris	FCN	96.15 ± 01.90	01.38 ± 01.16
	GAN	88.96 ± 08.98	04.57 ± 04.63
MICHE-I	FCN	80.49 ± 20.65	02.73 ± 02.76
	GAN	61.93 ± 24.97	10.95 ± 06.22
VIS	FCN	88.63 ± 09.15	02.47 ± 02.23
	GAN	72.15 ± 19.03	09.01 ± 05.54
All	**FCN**	**94.36 ± 09.90**	**01.26 ± 01.73**
	GAN	**86.62 ± 17.71**	**04.03 ± 05.28**

5.2.3.1.6 Visual & Qualitative Analysis

Here we perform a visual and qualitative analysis. First, in Fig. 5.16, we show poor and well-performed iris segmentation results obtained in each dataset by the FCN and GAN approaches. Some images were poorly segmented, thus explaining the high standard deviations obtained.

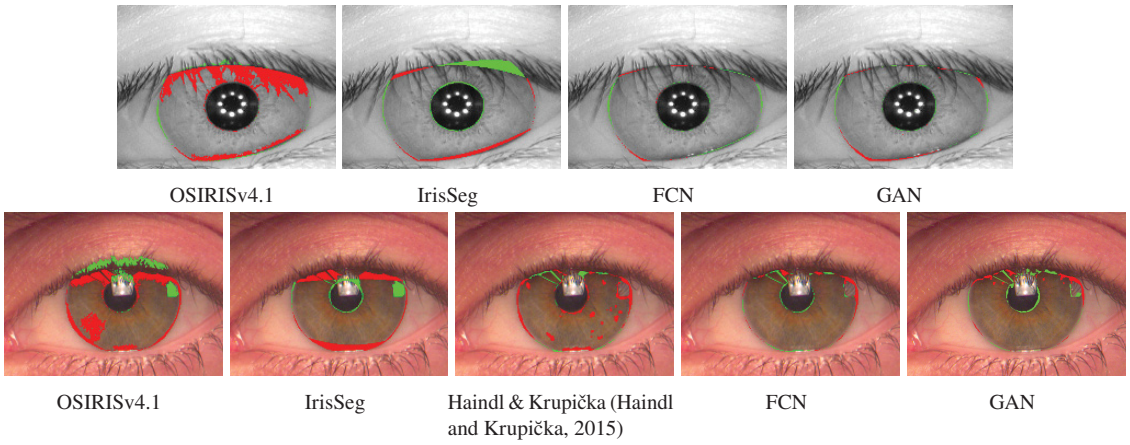
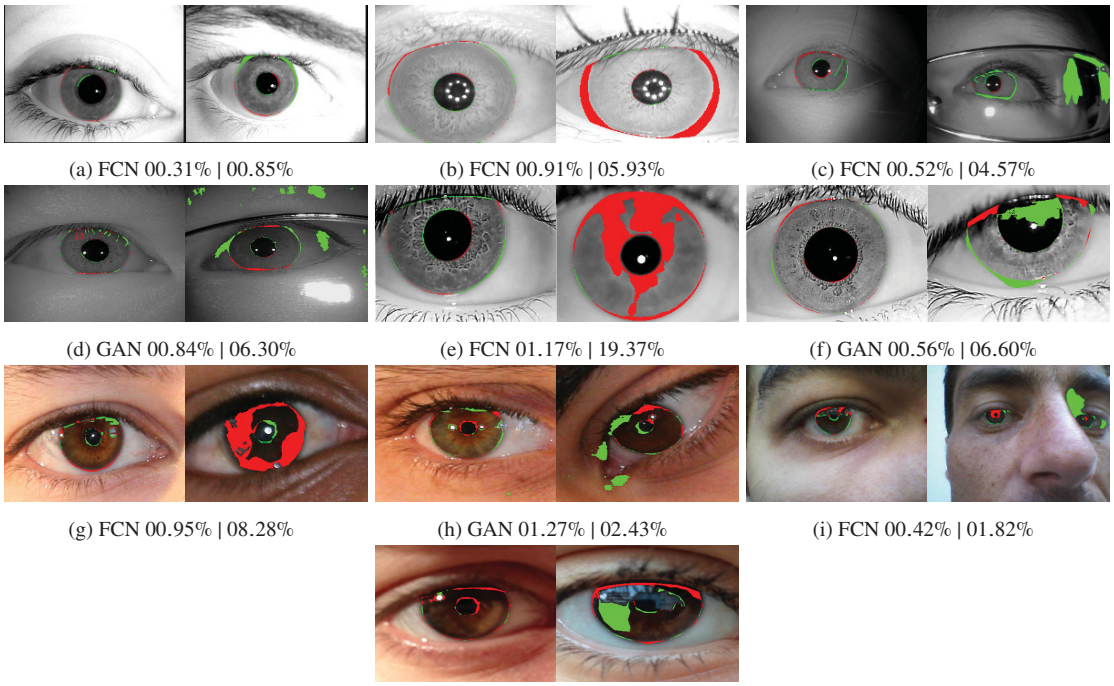


Figure 5.15: Qualitative results achieved by the FCN, GAN and baselines. Green and red pixels represent the FP and FN, respectively. The first and second rows correspond, respectively, to images from the CasiaI3 and CrEye-Iris datasets.

Then, in Fig. 5.15, we show iris segmentation performed by both the FCN and GAN approaches, as well as the baselines. We only show one image from each the CasiaI3 and CrEye-Iris datasets due to lack of space.

We particularly chose images where all methods perform fairly well and also where our methods performed better, which is the case in most situations. One can observe that our approach performed better in both NIR and VIS images.



(a) FCN 00.31% | 00.85% (b) FCN 00.91% | 05.93% (c) FCN 00.52% | 04.57%

(d) GAN 00.84% | 06.30% (e) FCN 01.17% | 19.37% (f) GAN 00.56% | 06.60%

(g) FCN 00.95% | 08.28% (h) GAN 01.27% | 02.43% (i) FCN 00.42% | 01.82%

(j) GAN 00.57% | 00.96%

Figure 5.16: FCN and GAN qualitative results: good (left) and bad (right) results based on the error E. Green and red pixels represent the FP and FN, respectively. (a)-(b) BioSec; (c)-(d) CasiaI3; (e)-(f) CasiaT4; (g)-(h) IITD Iris Image Database 1.0 (IITD-1); (i)-(j).

5.2.4 Final Remarks

Once we validate the hypothesis that it was or was not possible to segment the ORC with convolutional neural networks, we direct our efforts to investigated contextual based the simultaneous segmentation of the iris and the sclera regions, the proposed approaches and the obtained results can be seen in Section 5.3.

5.3 CONTEXT BASED OCULAR REGION COMPONENTS SEGMENTATION

As stated previously, the ocular region components can be employed as input to biometrics, once these components present a high level of differentiation between users (Wildes, 1997; Jain et al., 2004; Das et al., 2016, 2017; Delna et al., 2016). However, the images need to be submitted to a preprocessing stage to extract the RoI. The preprocessing stage has great importance in a biometric system, because if the RoI extraction is erroneously performed, the system effectiveness may be injured (patterns can be removed and/or introduced into the RoI) (Lucio et al., 2018; Rattani and Derakhshani, 2017).

Besides the high number of component-based preprocessing segmentation approaches, none of them works correctly in specific scenarios, such as in non-controlled environments. Trying to solve this many works were developed using machine learning approaches (Lucio et al., 2018; Bezerra et al., 2018; Jalilian et al., 2017; Liu et al., 2016a; Severo et al., 2018), mainly due the fact of deep learning have presented good results on many others correlated areas such as face recognition, natural language processing, speech recognition. However none of them evaluate the contextual information present in the images

The use of CNNs to extract the RoIs, taking into account context information is a promising approach, due to: (i) the available segmentation approaches do not work correctly in

some scenarios; (ii) the context in which the RoI extraction is performed is not used by currently available approaches; (iii) the capabilities of decision power presented by the deep learning approaches.

Thus, this section address the the main question addressed by this method is "Can we improve the RoI extractor in ocular region components taking into account the context information present in an image?" Aiming to answer this question, we propose the Ocular Region Context Network (ORCNet) as a new segmentation approach to simultaneous segment the sclera and iris using contextual information. The proposed approach combines the Context Encoding Network (EncNet) with the main contribution of this work, the Punish Context Loss (PC-Loss). The PC-Loss consists of punishing the segmentation *Losses* of a network by using a percentage difference value between the ground truth and the segmented masks. The proposed approach employs Biederman's semantic relationship concepts, in which we use three contexts (semantic, spatial, and scale) to evaluate the relationships of the objects in an image.

5.3.0.1 Proposed Approach

This section presents the proposed architecture addressing the ocular region components (iris and sclera) segmentation, taking into account the contextual information of the region.

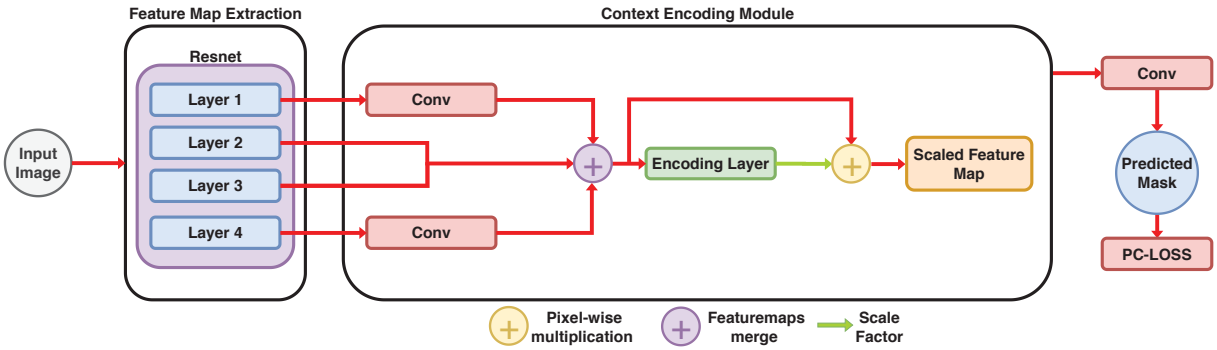


Figure 5.17: General overview of the proposed approach.

To solve the aforementioned segmentation problem, we propose the Ocular Region Context Network (ORCNet), a new architecture that combines the Context Encoding Network (EncNet) presented by Zhang et al. (2018) with the main contribution of this work: the Punish Context Loss (PC-Loss).

The Context Encoding Network (EncNet) was choice to compose the first stages of the proposed approach since the *context encoding module* present in this can understand the semantic relationships of the objects that make up an image. PC-Loss was created to supplement contextual information extracted by the EncNet. In the PC-Loss we have proposed 2 context coefficients values to evaluate the *spatial* and the *scale* relationships among the objects in a image. More details about the context coefficients are present in the nest paragraphs. In the EncNet, the extraction of feature maps is performed by using a ResNet based backbone, which was chosen because it does not suffer the effects of the vanishing gradient, thus it does not compromise the network overall performance. Once the feature maps are extracted using the ResNet backbone, they were used to feed the *Context Encoding Module*.

The *Context Encoding Module* is composed of an *Encoding Layer* (Zhang et al., 2017) which captures the encoded semantics, a fully connected layer, and a sigmoid as the activation function, which output scaling factors $\gamma = \delta(W_e)$, where W denotes the layer weights and δ

is the sigmoid function. Then, the module output is given by $Y = X \otimes \gamma$, where $\otimes$ denotes a channel-wise multiplication between the input feature maps X and scaling factors γ. The channel-wise multiplication is employed to emphasize or de-emphasize class-dependent feature maps.

Once the feature maps are processed by the EncNet, we employ the Punish Context Loss (PC-Loss). The PC-Loss consists of punishing the segmentation $Loss$ of the network by using a percentage value obtained, taking into account the Biederman's(Biederman, 1972) relationships, i.e.,

$$\text{PC-Loss} = \frac{\lambda + \rho}{2}, \tag{5.7}$$

where the parameter λ is the *Scale Context Coefficient* (a value that estimates the average difference in terms of *Jaccard Distance* from a class to all other classes on the addressed problem) of an image and it is defined as:

$$\lambda = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N-1}\sum_{\substack{j=1 \\ j\neq i}}^{N}\sqrt[2]{(\theta_i - \theta_j)^2}, \tag{5.8}$$

where N is the number of classes present in the evaluated problem, and the θ term is the *Jaccard Distance* obtained by subtracting the *Jaccard similarity coefficient* $J = \frac{gt_i \cap prd_i}{gt_i \cup prd_i}$ from 1, where gt_i and prd_i stand for the ground truth and predicted masks, respectively, i.e.,

$$\theta_i = 1 - J. \tag{5.9}$$

The ρ parameter is the ratio between the *Spatial Context Coefficient* δ from the ground truth and the predicted segmentation,

$$\rho = \frac{\delta(prd)}{\delta(gt)}. \tag{5.10}$$

where

$$\delta = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N-1}\sum_{\substack{j=1 \\ i\neq j}}^{N}\sqrt[2]{(C_i - C_j)^2}, \tag{5.11}$$

C the center of mass from an evaluated object. Considering that an object consists of n distinct points $x_1 ... x_n$, then the centroid (center of mass) is given by,

$$C = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{5.12}$$

The ORCNet implementation was made in the PyTorch framework, and to perform the training we adopted a learning rate of 10^{-4} combined with a momentum of 0.9 during 1000 epochs.

5.3.0.2 Experimental Protocol and Results

In this section are presented the baseline (see Section 5.3.0.2.1) , the datasets (see Section 5.3.0.2.2) , the evaluation protocol (see Section 5.3.0.2.3) and the results (see Section 5.3.0.2.4) obtained by using the Ocular Region Context Network (ORCNet).

5.3.0.2.1 Baselines

We selected three baseline frameworks described (and available) in the literature to compare to our approach: EncNet (Zhang et al., 2018), Fully Convolutional Network (FCN) (Long et al., 2015) and Generative Adversarial Network (GAN) (Goodfellow et al., 2014).

The FCN and the GAN segmentation approaches were chosen as baseline methods because they presented promising results on sclera and iris segmentation tasks (Bezerra et al., 2018; Lucio et al., 2018). The EncNet approach was chosen since the *Context Encoding Module* proposed in it encodes the contextual information in the image, thus outperforming the conventional segmentation approaches employed in many different research areas.

5.3.0.2.2 Datasets

To evaluate the performance of the Ocular Region Context Network (ORCNet) we Mobile Iris Challenge Evaluation I (MICHE-I) dataset. The dataset was chosen because it is composed of 3,191 non-processed images captured using three different devices that makes the work of segmenting the components of the ocular region even more challenging (see Table 5.15).

Table 5.15: Overview of the databases used in this work (Marsico et al., 2015).

Capture device	Images	Subjects	Resolution
Samsung Galaxy S4	1,297	92	2322×4128 and 1080×1920
Samsung Galaxy Tab 2	632	92	640×480
Apple iPhone 5	1,262	92	1536×2048 and 960×1280

5.3.0.2.3 Evaluation Protocol

The evaluation of an automatic detection approach is performed in a pixel-to-pixel comparison between the ground truth and the predicted segmentation mask. Therefore, we use the mean F-score, Intersection over Union (IoU) and Error Rate (ER) evaluation metrics.

In order to perform a fair evaluation and comparison of the proposed approaches, we divided each database into three subsets, being 40% of the images for training, 40% for testing and 20% for validation. We adopted this protocol (i.e., with a larger test set) to provide more samples for the sake of statistical significance.

5.3.0.2.4 Results and Discussions

Table 5.16 presents the mean results obtained in the performed experiments. This table is divided into three sections, one to each segmentation scenario (ALL (Iris and Sclera are considered as a unique region), Iris and Sclera). For each one of the scenarios, four segmentation approaches were evaluated, that is, EncNet, FCN, GAN, and our proposal ORCNet. Table 5.3.0.2.4 in turn presents the individual result for each MICHE-I database subset for the same scenario division presented in Table 5.16.

Table 5.16: Performance comparison among the baseline approaches and the proposed Ocular Region Context Network approach employing the MICHE-I dataset as input data. Intersection over Union is considered as the prior measure ranking methods.

Dataset	Approach	RoI	F-Score	ER	IoU	Dataset	Approach	RoI	F-Score	ER	IoU	Dataset	Approach	RoI	F-Score	ER	IoU
MICHE-GS4	EncNet ResNet-50	ALL	93.77 ± 5.31	0.34 ± 0.22	86.31 ± 11.81	MICHE-GT2	EncNet ResNet-50	ALL	93.57 ± 5.63	0.57 ± 0.39	83.85 ± 15.56	MICHE-IP5	EncNet ResNet-50	ALL	95.29 ± 3.96	0.25 ± 0.25	89.38 ± 12.57
	EncNet ResNet-101		93.40 ± 5.41	0.40 ± 0.24	83.80 ± 13.41		EncNet ResNet-101		94.03 ± 6.34	0.51 ± 0.39	85.38 ± 15.97		EncNet ResNet-101		95.25 ± 4.11	0.26 ± 0.24	88.51 ± 14.15
	EncNet ResNet-152		93.51 ± 4.95	0.37 ± 0.21	85.17 ± 12.82		EncNet ResNet-152		94.01 ± 5.83	0.54 ± 0.46	84.75 ± 16.98		EncNet ResNet-152		95.48 ± 5.83	0.25 ± 0.26	89.89 ± 11.97
	FCN		78.60 ± 13.19	1.22 ± 0.92	66.38 ± 15.36		FCN		89.56 ± 7.41	0.96 ± 1.04	81.02 ± 12.84		FCN		82.42 ± 12.57	1.06 ± 1.39	70.98 ± 16.70
	GAN		92.99 ± 7.20	0.37 ± 0.23	87.51 ± 9.75		GAN		86.31 ± 10.80	1.12 ± 0.68	76.72 ± 13.51		GAN		90.09 ± 7.68	0.53 ± 0.36	82.06 ± 11.75
	GAN PROCESSED		93.60 ± 6.13	0.34 ± 0.22	88.39 ± 8.53		GAN PROCESSED		88.25 ± 13.19	0.90 ± 0.65	80.31 ± 15.71		GAN PROCESSED		91.27 ± 8.34	0.46 ± 0.36	84.06 ± 11.46
	ORCNet ResNet-50		94.65 ± 3.65	0.28 ± 0.17	88.77 ± 9.26		ORCNet ResNet-50		95.19 ± 4.76	0.35 ± 0.25	89.92 ± 11.32		ORCNet ResNet-50		96.39 ± 4.24	0.19 ± 0.22	92.21 ± 11.11
	ORCNet ResNet-101		94.78 ± 3.47	0.28 ± 0.16	89.31 ± 7.47		ORCNet ResNet-101		95.40 ± 4.65	0.34 ± 0.25	90.30 ± 11.52		ORCNet ResNet-101		96.36 ± 4.77	0.19 ± 0.21	92.32 ± 10.76
	ORCNet ResNet-152		**95.09 ± 3.32**	**0.27 ± 0.16**	**89.91 ± 6.90**		ORCNet ResNet-152		**95.38 ± 4.92**	**0.34 ± 0.27**	**90.37 ± 11.32**		ORCNet ResNet-152		**96.53 ± 3.80**	**0.18 ± 0.23**	**92.62 ± 9.83**
	EncNet ResNet-50	Iris	94.67 ± 3.70	0.18 ± 0.13	88.36 ± 11.65		EncNet ResNet-50	Iris	94.17 ± 6.85	0.28 ± 0.24	86.55 ± 16.76		EncNet ResNet-50	Iris	95.91 ± 6.35	0.14 ± 0.21	90.28 ± 14.75
	EncNet ResNet-101		94.15 ± 4.82	0.20 ± 0.13	86.79 ± 13.04		EncNet ResNet-101		94.84 ± 5.86	0.26 ± 0.25	87.39 ± 17.10		EncNet ResNet-101		95.99 ± 6.40	0.14 ± 0.20	89.58 ± 16.15
	EncNet ResNet-152		94.66 ± 3.57	0.19 ± 0.13	87.51 ± 11.84		EncNet ResNet-152		94.79 ± 6.09	0.27 ± 0.26	87.10 ± 18.08		EncNet ResNet-152		95.74 ± 6.51	0.14 ± 0.20	90.34 ± 14.17
	FCN		69.10 ± 15.67	0.73 ± 0.55	52.98 ± 16.50		FCN		89.93 ± 7.08	0.56 ± 0.52	80.80 ± 14.41		FCN		83.16 ± 13.83	0.60 ± 0.79	71.95 ± 18.37
	GAN		80.12 ± 11.90	0.51 ± 0.25	68.13 ± 13.31		GAN		91.40 ± 9.59	0.44 ± 0.41	82.92 ± 17.39		GAN		93.53 ± 8.17	0.20 ± 0.27	87.39 ± 13.60
	GAN PROCESSED		82.34 ± 11.52	0.42 ± 0.22	71.25 ± 13.15		GAN PROCESSED		91.27 ± 10.32	0.45 ± 0.43	82.45 ± 18.63		GAN PROCESSED		93.81 ± 6.79	0.21 ± 0.26	87.07 ± 14.31
	ORCNet ResNet-50		94.81 ± 3.28	0.16 ± 0.11	89.39 ± 9.53		ORCNet ResNet-50		95.50 ± 4.68	0.21 ± 0.21	90.14 ± 13.54		ORCNet ResNet-50		96.55 ± 6.48	0.12 ± 0.21	91.74 ± 14.61
	ORCNet ResNet-101		94.98 ± 4.08	0.16 ± 0.10	89.98 ± 8.15		ORCNet ResNet-101		95.75 ± 5.27	0.20 ± 0.25	90.56 ± 13.92		ORCNet ResNet-101		96.38 ± 7.03	0.12 ± 0.20	91.93 ± 13.83
	ORCNet ResNet-152		**95.33 ± 2.73**	**0.15 ± 0.09**	**90.62 ± 7.58**		ORCNet ResNet-152		**95.96 ± 4.86**	**0.19 ± 0.20**	**90.75 ± 13.74**		ORCNet ResNet-152		**96.45 ± 7.30**	**0.12 ± 0.20**	**92.14 ± 13.94**
	EncNet ResNet-50	Sclera	85.52 ± 8.91	0.30 ± 0.21	73.09 ± 15.23		EncNet ResNet-50	Sclera	83.04 ± 13.10	0.48 ± 0.31	68.98 ± 18.46		EncNet ResNet-50	Sclera	88.67 ± 08.43	0.24 ± 0.21	77.21 ± 17.12
	EncNet ResNet-101		83.67 ± 9.64	0.33 ± 0.11	69.70 ± 16.80		EncNet ResNet-101		84.67 ± 12.04	0.45 ± 0.30	71.40 ± 18.43		EncNet ResNet-101		88.45 ± 7.70	0.25 ± 0.21	76.34 ± 18.43
	EncNet ResNet-152		85.17 ± 9.53	0.32 ± 0.19	70.90 ± 17.86		EncNet ResNet-152		83.81 ± 13.03	0.47 ± 0.37	70.37 ± 19.44		EncNet ResNet-152		88.74 ± 8.19	0.24 ± 0.23	77.35 ± 16.49
	FCN		69.10 ± 15.67	0.73 ± 0.55	52.98 ± 16.50		FCN		76.58 ± 14.07	0.79 ± 0.68	62.64 ± 16.75		FCN		71.54 ± 16.74	0.69 ± 0.77	56.44 ± 18.73
	GAN		80.12 ± 11.90	0.51 ± 0.25	68.13 ± 13.31		GAN		77.92 ± 14.83	0.81 ± 0.50	64.97 ± 16.88		GAN		83.00 ± 12.77	0.41 ± 0.34	71.92 ± 15.45
	GAN PROCESSED		82.34 ± 11.52	0.42 ± 0.22	71.25 ± 13.15		GAN PROCESSED		79.28 ± 14.41	0.74 ± 0.45	66.34 ± 16.90		GAN PROCESSED		84.54 ± 10.63	0.38 ± 0.32	73.18 ± 14.77
	ORCNet ResNet-50		88.27 ± 6.75	0.25 ± 0.16	78.42 ± 11.60		ORCNet ResNet-50		88.37 ± 10.07	0.32 ± 0.22	79.51 ± 14.49		ORCNet ResNet-50		91.62 ± 5.79	0.19 ± 0.18	82.49 ± 14.75
	ORCNet ResNet-101		88.44 ± 5.77	0.25 ± 0.17	79.02 ± 9.81		ORCNet ResNet-101		88.44 ± 10.76	0.32 ± 0.26	79.63 ± 14.84		ORCNet ResNet-101		91.49 ± 6.30	0.19 ± 0.18	82.56 ± 14.67
	ORCNet ResNet-152		**88.57 ± 5.60**	**0.24 ± 0.17**	**79.51 ± 9.41**		ORCNet ResNet-152		**88.85 ± 10.03**	**0.32 ± 0.24**	**79.68 ± 15.40**		ORCNet ResNet-152		**91.93 ± 5.27**	**0.18 ± 0.19**	**83.51 ± 13.71**

Yet, regarding the employed segmentation approaches, it is necessary to highlight that for the EncNet, ORCNet, and GAN approaches, more than one segmentation approach was employed. For the EncNet and ORCNet three segmentation variants were proposed, being the ResNet backbone (ResNet-50, ResNet-101 and ResNet-152) the difference among them. For the GAN approach, two different approaches were employed, one based on the method proposed by (Goodfellow et al., 2014) and another one in which small noise particles in the output images were removed by mathematical morphology operations. By analyzing the obtained results presented in Table 5.16, we observed that in all of the segmentation scenarios —ALL(Iris + Sclera), Iris and Sclera—, the ORCNet approach presented the best results in terms of $F - Score$, Intersection over Union (IoU), and Error Rate (ER). Also, it is important to observe that independently from the used ResNet backbone, the proposed ORCNet approach outperforms all the baselines evaluated in this work.

Once a general analysis of the performance obtained with the ORCNet was presented, we can individually discuss the results obtained in each of the scenarios evaluated using the protocol proposed in Section 5.3.0.2.3. In ALL segmentation scenario, we observed that the mean $F - Score$ obtained by using the ORCNet ResNet-152 is 95.67%, outperforming the best baseline (EncNet ResNet-152) score by 1.42%. In terms of ER the obtained score outperforms the best baseline by 33%, reducing the error rate of 0.39% to 0.26%. Finally, by analyzing the IoU, we observed that in the ALL segmentation scenario, the proposed approach achieved a score of 90.97%, outperforming the best baseline result by 5.04%.

The behaviour presented in the ALL (iris + sclera) segmentation scenario can also be seen in the iris and sclera individual segmentation tasks, since the best results were obtained by using the ORCNet. In the Iris segmentation the mean F-Score, ER and IoU obtained values are 95.91%, 0.15% and 91.17% outperforming the best baseline results by 0.89%, 25% and 3.22% respectively. And, in the sclera segmentation the mean F-Score, ER and IoU obtained values are 89.78%, 0.24% and 80.90%, outperforming the best results by 4.50%, 26.47% and 11.01%, respectively.

In addition to presenting better results in relation to the baselines, we also found that the standard deviation presented by the results was smaller when the context based segmentation was employed. In this way it is possible to state that the ORCNet presented most stable segmentation of the Region of Interest (RoI), since it showed less variability in the segmentation masks. Finally, it is important to highlight that regardless of the backbone used as feature maps extractor in ORCNet, the proposed approach outperforms all the baselines evaluated.

5.3.1 Final Remarks

This Chapter described and detailed the proposed methods, experimental protocols, and databases employed to evaluate the raised hypothesis. We started by our method to segment the sclera by combing a state-of-the-art segmentation approach with a -based object detection as a preprocessing step. Then, we described a CNN based method to simultaneous detect the periocular region. Additionally, we investigate the use of Generative Adversarial Network (GAN) and Fully Convolutional Network (FCN) the iris and the sclera in a independent way. Finally, we presented our new context-based segmentation approach to segment the periocular region components (sclera and iris). In the next chapter are present the conclusion about the use experiments performed in this work.

6 CONCLUSION

In this study, we developed a deep learning approach to segment ORCs. Considering the hypothesis that it is possible to achieve state-of-the-art results by employing a deep learning context-based approach to periocular region component RoI extraction, we suggested some interesting approaches.

To explore objective 1 presented in Section 1.5, we applied the CNN in the detection of the iris. To validate objective 1, we compared the HOG – a classical approach to detect the targets – against the YOLO approach. Considering the final scores and frame rate, YOLO had better results.

After performing iris detection in an iris-only evaluation, we proceed to the second objective presented in Section 1.5: to validate the use of a CNN for simultaneous detection of the iris and the ocular region. In this scenario, we annotated more than 170,000 images using a coarse-image annotation approach. We drew two BBOX for each image, one to bound the iris and another one to border the ocular region.

Once we defined the RoI and determine how to annotate it for the simultaneous detection task, we evaluated different detection approaches. In this new detection scenario, we compared the YOLO-based approach to the "Faster-R-CNN + FPN" concepts. After carrying out all the proposed experiments and assessing the results, the Faster R-CNN + FPN approach presented slightly better results than the YOLO-based approach. However, the YOLO-based approach showed better results in frame rate terms. We also verified that using simultaneous detection of the iris and ocular region achieves better general accuracy than the iris-only detection. These results leads us to believe that more than one target of interest is positive in a detection task as the model can understand some contextual information.

Since we were able to detect ORCs using a neural network, we found methods to segment these regions: sclera and iris. The first step to achieving this third objective of this study was to investigate the feasibility of employing CNN approaches in sclera segmentation. To validate this point, we proposed two new approaches: based on FCN and based on GAN. Sclera segmentation was possible running CNN, to date, both approaches, GAN and FCN obtaining state-of-the-art baseline considering the F-score on MICHE-I and UBIRIS.v2 image subsets.

The approach based on an FCN for sclera segmentation was submitted to "SSBC 2020: Sclera Segmentation Benchmarking Competition in the Mobile Environment" and was ranked 2nd place. After winning 2nd place, we were invited by the SSBC leader to a collaborative work in other subsets of visual images. Consequently, we won 1st place in another competition and is currently under review in the Transactions on Information Forensics and Security journal.

In addition to sclera segmentation experiments, we evaluated the segmentation of the iris using both FCN and GAN approaches. Iris segmentation using both methods yielded formidable results, hitting the state-of-the-art baseline again. Considering all the presented results on iris and sclera segmentation, we successfully accomplished the third objective proposed.

After the ORC detection and segmentation tasks, we aimed to validate the most interesting objective of this study: introducing an unprecedented approach to sclera and iris segmentation that can understand Biederman's contextual relationships (semantic, scale, and spatial components), currently called ORCNet. This approach combines EncNet with PC-Loss, our great contribution to ORC segmentation fields. By using EncNet, the proposed approach can perform deep comprehension of semantic relationships present in the images since the context encoding module evaluates the likelihood of a RoI being present or absent in an image.

Meanwhile, PC-Loss goes through scale and spatial contextual information present in the images. All cases where this proposed approach was applied had better results, reaching a gold standard in this field.

Supported by our experiments and results, we can affirm that a context-based OCR segmentation approach can achieve impressive results in an uncontrolled environment. However, the approach needs improvement to address some complex limitations, such as facial and ocular region segmentation in the wild and on NIR images. Therefore, we suggest, as future work, evaluating the proposed domains along with the proposed method as a scene understanding scenario.

REFERENCES

Aginako, N., Martínez-Otzeta, J. M., Rodriguez, I., Lazkano, E., and Sierra, B. (2016). Machine learning approach to dissimilarity computation: Iris matching. In *23rd ICPR*, pages 170–175.

Ahuja, K., Islam, R., Barbhuiya, F. A., and Dey, K. (2017). Convolutional neural networks for ocular smartphone-based biometrics. *Pattern Recognition Letters*, 91.

Al-Waisy, A. S., Qahwaji, R., Ipson, S., Al-Fahdawi, S., and Nagem, T. A. M. (2018). A multi-biometric iris recognition system based on a deep learning approach. *Pattern Analysis and Applications*, 21.

Algashaam, F. M., Nguyen, K., Alkanhal, M., Chandran, V., Boles, W., and Banks, J. (2017). Multispectral Periocular Classification With Multimodal Compact Multi-Linear Pooling. *IEEE Access*, 5:14572–14578.

Ali, N. and Zafar, B. (2018). MSRC-v2 image dataset.

Alvarez-Betancourt, Y. and Garcia-Silvente, M. (2010). A fast iris location based on aggregating gradient approximation using qma-owa operator. In *FUZZ-IEEE*, pages 1–8.

Arora, S. S., Vatsa, M., Singh, R., and Jain, A. (2012). Iris recognition under alcohol influence: A preliminary study. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 336–341. IEEE.

Audebert, N., Le Saux, B., and Lefèvre, S. (2017). Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In *Computer Vision – ACCV 2016*, pages 180–196.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39.

Baker, S. E., Bowyer, K. W., Flynn, P. J., and Phillips, P. J. (2013). Template Aging in Iris Biometrics. In Burge, M. J. and Bowyer, K. W., editors, *Handbook of Iris Recognition*, pages 205–218. Springer London, London.

Baker, S. E., Hentz, A., Bowyer, K. W., and Flynn, P. J. (2010). Degradation of iris recognition performance due to non-cosmetic prescription contact lenses. *Computer Vision and Image Understanding*, 114(9):1030–1044.

Bar, M. and Ullman, S. (1996). Spatial context in recognition. *Perception*, 25.

Barlow, H. B. (1989). Unsupervised learning. *Neural computation*, 1(3):295–311.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bezerra, C. S., Laroca, R., Lucio, D. R., Severo, E., Oliveira, L. F., Britto, A. S., and Menotti, D. (2018). Robust iris segmentation based on fully convolutional networks and generative adversarial networks. In *31st SIBGRAPI*, pages 281–288.

Biederman, I. (1972). Perceiving real-world scenes. *Science*, 177.

Bolle, R. M., Connell, J. H., Pankanti, S., Ratha, N. K., and Senior, A. W. (2004). *Guide to Biometrics*. Springer Publishing Company, Incorporated.

Bolme, D. S., Draper, B. A., and Beveridge, J. R. (2009). Average of synthetic exact filters. In *IEEE CVPR*, pages 2105–2112.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152.

Bowyer, K. W., Hollingsworth, K., and Flynn, P. J. (2008). Image understanding for iris biometrics: A survey. *Computer Vision and Image Understanding*, 110.

Cai, W., Zhang, B., and Wang, B. (2021). Scale-aware anchor-free object detection via curriculum learning for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:9946–9958.

Carbonetto, P., De Freitas, N., and Barnard, K. (2004). A statistical model for general contextual object recognition. In *European conference on computer vision*, pages 350–362. Springer.

CASIA (2002). Casia version 1 database.

CASIA (2004). Casia version 2 database.

CASIA (2010). Casia version 4 database.

Chang, K. I., Bowyer, K. W., Flynn, P. J., and Chen, X. (2004). Multi-biometrics using facial appearance, shape and temperature. In *IEEE FG*, pages 43–48.

Cheng, G., Han, J., Zhou, P., and Guo, L. (2014). Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98:119–132.

Cui, W. et al. (2012). A rapid iris location algorithm based on embedded. In *CSIP*, pages 233–236.

Cunningham, P., Cord, M., and Delany, S. J. (2008). Supervised learning. In *Machine learning techniques for multimedia*, pages 21–49. Springer.

Czajka, A. (2013). Database of Iris Printouts and its Application : Development of Liveness Detection Method for Iris Recognition. *MMAR, 18th International Conference*, pages 28–33.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893 vol. 1.

Das, A., Pal, U., Ballester, M. A. F., and Blumenstein, M. (2013). Sclera recognition using dense-SIFT. In *13th ISDA*, pages 74–79.

Das, A., Pal, U., Ferrer, M. A., and Blumenstein, M. (2015). SSBC 2015: Sclera segmentation benchmarking competition. In *2015 IEEE 7th Int. Conf. on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6.

Das, A., Pal, U., Ferrer, M. A., and Blumenstein, M. (2016). SSRBC. In *2016 Int. Conf. on Biometrics (ICB)*, pages 1–6.

Das, A., Pal, U., Ferrer, M. A., Blumenstein, M., Štepec, D., Rot, P., Emeršič, Ž., Peer, P., Struc, Š., Kumar, S. V. A., and Harish, B. S. (2017). SSERBC 2017: Sclera segmentation and eye recognition benchmarking competition. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 742–747.

Daugman, J. (2004). How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):21–30.

Daugman, J. (2007). New methods in iris recognition. *IEEE TSMC, Part B*, 37(5):1167–1175.

Daugman, J. G. (1993). High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1148–1161.

Delna, K. V., Sneha, K. A., and Aneesh, R. P. (2016). Sclera vein identification in real time using single board computer. In *2016 Int. Conf. on Next Generation Intelligent Systems (ICNGIS)*, pages 1–5.

Deshpande, A., Dubey, S., Shaligram, H., Potnis, A., and Chavan, S. (2014). Iris recognition system using block based approach with DWT and DCT. In *IEEE Anual India Conference*, pages 1–5.

Ding, Y., Tao, Q., Wang, L., Li, D., and Zhang, M. (2018). Image-based localisation using shared-information double stream hourglass networks. *Electronics Letters*, 54(8):496–498.

Doyle, J. and Bowyer, K. (2014). Notre dame image database for contact lens detection in iris recognition.

Doyle, J. S. and Bowyer, K. W. (2015a). Robust detection of textured contact lenses in iris recognition using BSIF. *IEEE Access*, 3:1672–1683.

Doyle, J. S. and Bowyer, K. W. (2015b). Robust Detection of Textured Contact Lenses in Iris Recognition Using BSIF. *IEEE Access*, 3:1672–1683.

Doyle, J. S., Bowyer, K. W., and Flynn, P. J. (2013). Variation in accuracy of textured contact lens detection based on sensor and lens pattern. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE.

Dumoulin, V. and Visin, F. (2016). A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2007). The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Fenker, S. P. and Bowyer, K. W. (2012). Analysis of template aging in iris biometrics. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 45–51. IEEE.

Fierrez, J., Ortega-Garcia, J., Torre Toledano, D., and Gonzalez-Rodriguez, J. (2007). Biosec baseline corpus: A multimodal biometric database. *Pattern Recognition*, 40(4):1389–1392.

Fink, M. and Perona, P. (2003). Mutual boosting for contextual inference. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 1515–1522, Cambridge, MA, USA. MIT Press.

Fischler, M. A. and Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22(1):67–92.

Franchi, G., Angulo, J., and Sejdinović, D. (2016). Hyperspectral image classification with support vector machines on kernel distribution embeddings. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1898–1902.

Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer.

Galleguillos, C., Rabinovich, A., and Belongie, S. (2008). Object categorization using co-occurrence, location and appearance. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Gangwar, A., Joshi, A., Singh, A., Alonso-Fernandez, F., and Bigun, J. (2016a). IrisSeg: A fast and robust iris segmentation framework for non-ideal iris images. In *Int. Conf. on Biometrics (ICB)*, pages 1–8.

Gangwar, A., Joshi, A., Singh, A., Alonso-Fernandez, F., and Bigun, J. (2016b). IrisSeg: A fast and robust iris segmentation framework for non-ideal iris images. In *International Conference on Biometrics (ICB)*, pages 1–8.

Girshick, R. (2015). Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 580–587, Washington, DC, USA. IEEE Computer Society.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *NIPS*, pages 2672–2680. Curran Associates, Inc.

Gupta, P., Behera, S., Vatsa, M., and Singh, R. (2014). On Iris Spoofing Using Print Attack. In *2014 22nd International Conference on Pattern Recognition*, pages 1681–1686. IEEE.

Haindl, M. and Krupička, M. (2015). Unsupervised detection of non-iris occlusions. *Pattern Recognition Letters*, 57:60–65.

Hanson, A. (1978). Visions: A computer system for interpreting scenes. *Computer vision systems*.

He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017). Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*.

He, X., Zemel, R. S., and Carreira-Perpiñán, M. A. (2004). Multiscale conditional random fields for image labeling. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR'04, pages 695–703, Washington, DC, USA. IEEE Computer Society.

Henry, C., Azimi, S. M., and Merkle, N. (2018). Road segmentation in SAR satellite images with deep fully-convolutional neural networks. *CoRR*, abs/1802.01445.

Hoiem, D., Efros, A. A., and Hebert, M. (2008). Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15.

Hosseini, M. S., Araabi, B. N., and Soltanian-Zadeh, H. (2010). Pigment melanin: Pattern for iris recognition. *IEEE Transactions on Instrumentation and Measurement*, 59(4):792–804.

IRISKING (2017). Irisking.

Isola, P., Zhu, J., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004.

Jain, A. K., Bolle, R., and Pankanti, S. (1998). *Biometrics, Personal Identification in Networked Society: Personal Identification in Networked Society*. Kluwer Academic Publishers, Norwell, MA, USA.

Jain, A. K., Nandakumar, K., and Ross, A. (2016). 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 79:80–105.

Jain, A. K., Ross, A., and Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20.

Jalilian, E., Uhl, A., and Kwitt, R. (2017). Domain adaptation for cnn based iris segmentation. In *Int. Conf. of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6.

Jillela, R. and Ross, A. A. (2015). Segmenting iris images in the visible spectrum with applications in mobile biometrics. *Pattern Recognition Letters*, 57:4–16.

Juefei-Xu, F. and Savvides, M. (2012). Unconstrained periocular biometric acquisition and recognition using COTS PTZ camera for uncooperative and non-cooperative subjects. In *IEEE Workshop on the Applications of Computer Vision (WACV)*, pages 201–208.

Kim, D., Jung, Y., Toh, K.-A., Son, B., and Kim, J. (2016). An empirical study on iris recognition in a mobile phone. *Expert Systems with Applications*, 54:328 – 339.

Kimura., G., Lucio., D., Britto Jr.., A., and Menotti., D. (2020). Cnn hyperparameter tuning applied to iris liveness detection. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 428–434. INSTICC, SciTePress.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Ko, K. E. and Sim, K. B. (2017). Real-time object entity detection system for smart surveillance application. *Electronics Letters*, 53(19):1304–1306.

Kohli, N., Yadav, D., Vatsa, M., and Singh, R. (2013). Revisiting iris recognition with color cosmetic contact lenses. In *2013 International Conference on Biometrics (ICB)*, pages 1–7.

Kohli, N., Yadav, D., Vatsa, M., Singh, R., and Noore, A. (2016). Detecting Medley of Iris Spoofing Attacks using DESIST Naman Kohli. In *8th IEEE International Conference on Biometrics: Theory, Applications, and Systems*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kumar, S. and Hebert, M. (2005). A hierarchical field framework for unified context-based classification. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1284–1291 Vol. 2.

Kumar, S., Loui, A. C., and Hebert, M. (2003). An observation-constrained generative approach for probabilistic classification of image regions. *Image and Vision Computing*, 21(1):87–97.

Laroca, R., Cardoso, E. V., Lucio, D. R., Estevam, V., and Menotti, D. (2022). On the cross-dataset generalization for license plate recognition. *arXiv preprint arXiv:2201.00267*.

Laroca, R., Severo, E., Zanlorensi, L. A., Oliveira, L. S., Gonçalves, G. R., Schwartz, W. R., and Menotti, D. (2018). A robust real-time automatic license plate recognition based on the yolo detector. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10.

Laroca, R., Zanlorensi, L. A., Gonçalves, G. R., Todt, E., Schwartz, W. R., and Menotti, D. (2021). An efficient and layout-independent automatic license plate recognition system based on the yolo detector. *IET Intelligent Transport Systems*, 15(4):483–503.

Läthén, G., Andersson, T., Lenz, R., and Borga, M. (2009). Momentum based optimization methods for level set segmentation. In Tai, X.-C., Mørken, K., Lysaker, M., and Lie, K.-A., editors, *Scale Space and Variational Methods in Computer Vision*, pages 124–136, Berlin, Heidelberg. Springer Berlin Heidelberg.

Le, T. H. N., Prabhu, U., and Savvides, M. (2014). A novel eyebrow segmentation and eyebrow shape-based identification. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer.

Leng, J., Liu, Y., Zhang, T., and Quan, P. (2018a). Context learning network for object detection. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 667–673.

Leng, J., Liu, Y., Zhang, T., Quan, P., and Cui, Z. (2018b). Context-aware u-net for biomedical image segmentation. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2535–2538.

Li, K., Wan, G., Cheng, G., Meng, L., and Han, J. (2020). Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *CoRR*, abs/1312.4400.

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944.

Liu, N., Li, H., Zhang, M., Liu, J., Sun, Z., and Tan, T. (2016a). Accurate iris segmentation in non-cooperative environments using fully convolutional networks. In *Int. Conf. on Biometrics*, pages 1–8.

Liu, N., Zhang, M., Li, H., Sun, Z., and Tan, T. (2016b). Deepiris: Learning pairwise filter bank for heterogeneous iris verification. *Pattern Recognition Letters*, 82:154 – 161. An insight on eye biometrics.

Liu, X., Bowyer, K. W., and Flynn, P. J. (2005). Experiments with an improved iris segmentation algorithm. In *IEEE AutoID'05*, pages 118–123.

Liu, Y., Li, Q., Yuan, Y., and Wang, Q. (2022). Single-shot balanced detector for geospatial object detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2529–2533. IEEE.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440.

Long, Y., Gong, Y., Xiao, Z., and Liu, Q. (2017). Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498.

Luc, P., Couprie, C., Chintala, S., and Verbeek, J. (2016). Semantic segmentation using adversarial networks. *CoRR*, abs/1611.08408.

Lucio, D. R., Laroca, R., Severo, E., Britto Jr., A. S., and Menotti, D. (2018). Fully convolutional networks and generative adversarial networks applied to sclera segmentation. In *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7.

Lucio, D. R., Laroca, R., Zanlorensi, L. A., Moreira, G., and Menotti, D. (2019). Simultaneous iris and periocular region detection using coarse annotations. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 178–185.

Luz, E., Moreira, G., Junior, L. A. Z., and Menotti, D. (2018). Deep periocular representation aiming video surveillance. *Pattern Recognition Letters*, 114:2 – 12. Data Representation and Representation Learning for Video Analysis.

Mahalingam, G., Ricanek, K., and Albert, A. M. (2014). Investigating the periocular-based face recognition across gender transformation. *IEEE Transactions on Information Forensics and Security*, 9(12):2180–2192.

Marsico, M., Nappi, M., and Proença, H. (2017). Results from MICHE II – Mobile Iris CHallenge Evaluation II. *Pattern Recognition Letters*, 91:3–10.

Marsico, M., Nappi, M., Riccio, D., and Wechsler, H. (2015). Mobile iris challenge evaluation (MICHE)-I, biometric iris dataset and protocols. *Pattern Recognition Letters*, 57:17–23.

Mashudi, N. A., Ahmad, N., and Noor, N. M. (2021). Dynamic u-net using residual network for iris segmentation. In *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 117–121.

Menotti, D., Chiachia, G., Pinto, A., Schwartz, W. R., Pedrini, H., Falcao, A. X., and Rocha, A. (2015). Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security*, 10(4):864–879.

Nalla, P. R. and Kumar, A. (2017a). Toward More Accurate Iris Recognition Using Cross-Spectral Matching. *IEEE Transactions on Image Processing*, 26(1):208–221.

Nalla, P. R. and Kumar, A. (2017b). Toward More Accurate Iris Recognition Using Cross-Spectral Matching. *IEEE Transactions on Image Processing*, 26(1):208–221.

Naqvi, R. A. and Loh, W.-K. (2019). Sclera-net: Accurate sclera segmentation in various sensor images based on residual encoder and decoder network. *IEEE Access*, 7:98208–98227.

Nguyen, K., Fookes, C., Ross, A., and Sridharan, S. (2018). Iris recognition with off-the-shelf cnn features: A deep learning perspective. *IEEE Access*, 6:18848–18855.

Nigam, I., Vatsa, M., and Singh, R. (2015). Ocular biometrics: A survey of modalities and fusion approaches. *Information Fusion*, 26:1–35.

Osorio-Roig, D., Rathgeb, C., Gomez-Barrero, M., Morales-González, A., Garea-Llano, E., and Busch, C. (2018). Visible wavelength iris segmentation: A multi-class approach using fully convolutional neuronal networks. In Brömme, A., Busch, C., Dantcheva, A., Rathgeb, C., and Uhl, A., editors, *BIOSIG 2018 - Proceedings of the 17th International Conference of the Biometrics Special Interest Group*, Bonn. Köllen Druck+Verlag GmbH.

Othman, N., Dorizzi, B., and Garcia-Salicetti, S. (2016a). OSIRIS: An open source iris recognition software. *Pat. Rec. Letters*, 82:124–131.

Othman, N., Dorizzi, B., and Garcia-Salicetti, S. (2016b). OSIRIS: An open source iris recognition software. *Pat. Rec. Letters*, 82:124 – 131.

Ouabida, E., Essadique, A., and Bouzid, A. (2017). Vander lugt correlator based active contours for iris segmentation and tracking. *Expert Systems with Applications*, 71:383–395.

Padole, C. N. and Proença, H. (2012). Periocular recognition: Analysis of performance degradation factors. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 439–445. IEEE.

Park, U., Jillela, R. R., Ross, A., and Jain, A. K. (2011). Periocular biometrics in the visible spectrum. *IEEE Transactions on Information Forensics and Security*, 6(1):96–106.

Peláez, J. and Doña, J. (2006). A majority model in group decision making using QMA–OWA operators. *International Journal of Intelligent Systems*, 21(2):193–208.

Phillips, P. J., Bowyer, K. W., Flynn, P. J., Liu, X., and Scruggs, W. T. (2008). The iris challenge evaluation 2005. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–8.

Phillips, P. J., Scruggs, W. T., O'Toole, A. J., Flynn, P. J., Bowyer, K. W., Schott, C. L., and Sharpe, M. (2010). FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):831–846.

Podder, P., Khan, T. Z., Khan, M. H., Rahman, M. M., Ahmed, R., and Rahman, M. S. (2015). An efficient iris segmentation model based on eyelids and eyelashes detection in iris recognition system. In *Int. Conf. on Computer Communication and Informatics*, pages 1–7.

Proença, H. and Alexandre, L. A. (2005). UBIRIS: A noisy iris image database. In *Image Analysis and Processing – ICIAP*, pages 970–977.

Proença, H. and Alexandre, L. A. (2012). Toward covert iris biometric recognition: Experimental results from the NICE contests. *IEEE Trans. on Information Forensics and Security*, 7(2):798–808.

Proenca, H., Filipe, S., Santos, R., Oliveira, J., and Alexandre, L. A. (2010). The UBIRIS.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1529–1535.

Proença, H. and Alexandre, L. A. (2006). Iris segmentation methodology for non-cooperative recognition. *IEE Proceedings - Vision, Image and Signal Processing*, 153(2):199–205.

Proença, H. and Neves, J. C. (2017). Irina: Iris recognition (even) in inaccurately segmented data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6747–6756.

Proença, H. and Neves, J. C. (2018). Deep-prwis: Periocular recognition without the iris and sclera using deep learning frameworks. *IEEE Transactions on Information Forensics and Security*, 13(4):888–896.

Proença, H., Neves, J. C., and Santos, G. (2014). Segmenting the periocular region using a hierarchical graphical model fed by texture/shape information and geometrical constraints. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–7.

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., and Belongie, S. (2007). Objects in context. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Radu, P., Ferryman, J., and Wild, P. (2015). A robust sclera segmentation algorithm. In *2015 IEEE 7th Int. Conf. on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6.

Rattani, A. and Derakhshani, R. (2017). Ocular biometrics in the visible spectrum: A survey. *Image and Vision Computing*, 59:1–16.

Rattani, A., Derakhshani, R., Saripalle, S. K., and Gottemukkula, V. (2016). ICIP 2016 competition on mobile ocular biometric recognition. In *2016 IEEE International Conference on Image Processing (ICIP)*, volume 2016-Augus, pages 320–324. IEEE.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016a). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

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

Redmon, J. and Farhadi, A. (2017a). Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271.

Redmon, J. and Farhadi, A. (2017b). YOLO9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 91–99, Cambridge, MA, USA. MIT Press.

Rodríguez, J. L. G. and Rubio, Y. D. (2005). A new method for iris pupil contour delimitation and its application in iris texture parameter estimation. In Sanfeliu, A. and Cortés, M. L., editors, *Progress in Pattern Recognition, Image Analysis and Applications*, pages 631–641, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rot, P., Emeršič, Ž., Struc, V., and Peer, P. (2018). Deep multi-class eye segmentation for ocular biometrics. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–8.

Roth, H. R., Oda, H., Zhou, X., Shimizu, N., Yang, Y., Hayashi, Y., Oda, M., Fujiwara, M., Misawa, K., and Mori, K. (2018). An application of cascaded 3d fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66:90–99.

Ruiz, P., Mateos, J., Camps-Valls, G., Molina, R., and Katsaggelos, A. K. (2014). Bayesian active remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(4):2186–2196.

Ruiz-Albacete, V., Tome-Gonzalez, P., Alonso-Fernandez, F., Galbally, J., Fierrez, J., and Ortega-Garcia, J. (2008). Direct attacks using fake images in iris verification. *Lecture Notes in Computer Science*, 5372 LNCS:181–190.

Santos, G., Grancho, E., Bernardo, M. V., and Fiadeiro, P. T. (2015). Fusing iris and periocular information for cross-sensor recognition. *Pattern Recognition Letters*, 57:52–59.

Sequeira, A., Chen, L., Wild, P., Ferryman, J., Alonso-Fernandez, F., Raja, K. B., Raghavendra, R., Busch, C., and Bigun, J. (2016). Cross-Eyed - Cross-Spectral Iris/Periocular Recognition Database and Competition. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, volume P-260, pages 1–5. IEEE.

Sequeira, A. F., Monteiro, J. C., Rebelo, A., and Oliveira, H. P. (2014a). Mobbio: A multimodal database captured with a portable handheld device. In *2014 International conference on computer vision theory and applications (VISAPP)*, volume 3, pages 133–139. IEEE.

Sequeira, A. F., Murari, J., and Cardoso, J. S. (2014b). Iris liveness detection methods in mobile applications. In *2014 International Conference on Computer Vision Theory and Applications*, volume 3, pages 22–33.

Sequeira, A. F., Murari, J., and Cardoso, J. S. (2014c). Iris Liveness Detection Methods in Mobile Applications. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 3, pages 22–33.

Severo, E., Laroca, R., Bezerra, C. S., Zanlorensi, L. A., Weingaertner, D., Moreira, G., and Menotti, D. (2018). A benchmark for iris location and a deep learning detector evaluation. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Shah, S. and Ross, A. (2006). Generating Synthetic Irises by Feature Agglomeration. In *2006 International Conference on Image Processing*, pages 317–320. IEEE.

Shah, S. and Ross, A. (2009). Iris segmentation using geodesic active contours. *IEEE Transactions on Information Forensics and Security*, 4(4):824–836.

Sharma, A., Verma, S., Vatsa, M., and Singh, R. (2014). On cross spectral periocular recognition. *2014 IEEE International Conference on Image Processing, ICIP 2014*, pages 5007–5011.

Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2009). Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. J. Comput. Vision*, 81(1):2–23.

Silva, P., Luz, E., Silva, G., Moreira, G., Silva, R., Lucio, D., and Menotti, D. (2020). Covid-19 detection in ct images with deep learning: A voting-based scheme and cross-datasets analysis. *Informatics in medicine unlocked*, 20:100427.

Silva, P. H., Luz, E., Zanlorensi, L. A., Menotti, D., and Moreira, G. (2018). Multimodal feature level fusion based on particle swarm optimization with deep transfer learning. In *IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Strat, T. M. and Fischler, M. A. (1991). Context-based vision: recognizing objects using information from both 2 d and 3 d imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1050–1065.

Su, L., Wu, J., Li, Q., and Liu, Z. (2017). Iris location based on regional property and iterative searching. In *2017 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 1064–1068.

Szegedy, C., , , Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Tan, C. and Kumar, A. (2012). Human identification from at-a-distance images by simultaneously exploiting iris and periocular features. In *International Conference on Pattern Recognition*, pages 553–556.

Tan, C. W. and Kumar, A. (2013). Towards online iris and periocular recognition under relaxed imaging constraints. *IEEE Transactions on Image Processing*, 22(10):3751–3765.

Tan, T., He, Z., and Sun, Z. (2010). Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition. *Image and Vision Computing*, 28(2):223–230.

Tan, T. and Sun, Z. (2005). CASIA-IrisV3. *Chinese Academy of Sciences Institute of Automation, http://www.cbsr.ia.ac. cn/IrisDatabase.htm, Tech. Rep*.

Tapia, J. E., Perez, C. A., and Bowyer, K. W. (2016). Gender Classification From the Same Iris Code Used for Recognition. *IEEE Transactions on Information Forensics and Security*, 11(8):1760–1770.

Teichmann, M., Weber, M., Zoellner, M., et al. (2016a). Multinet: Real-time joint semantic reasoning for autonomous driving. *arXiv preprint arXiv:1612.07695*.

Teichmann, M., Weber, M., Zöllner, J. M., Cipolla, R., and Urtasun, R. (2016b). Multinet: Real-time joint semantic reasoning for autonomous driving. *CoRR*, abs/1612.07695.

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191.

Torralba, A., Murphy, K. P., and Freeman, W. T. (2004). Contextual models for object detection using boosted random fields. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pages 1401–1408, Cambridge, MA, USA. MIT Press.

Tu, Z. and Bai, X. (2010). Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1744–1757.

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., and Smeulders, A. W. M. (2013). Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171.

Vargas, A. C. G., Paes, A., and Vasconcelos, C. N. (2016). Um estudo sobre redes neurais convolucionais e sua aplicação em detecção de pedestres. In *Proceedings of the xxix conference on graphics, patterns and images*, volume 1.

Verbeek, J. and Triggs, B. (2007). Scene segmentation with conditional random fields learned from partially labeled images. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 1553–1560, USA. Curran Associates Inc.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–511–I–518 vol.1.

Vitek, M., Das, A., Pourcenoux, Y., Missler, A., Paumier, C., Das, S., De Ghosh, I., Lucio, D. R., Zanlorensi, L. A., Menotti, D., et al. (2020). Ssbc 2020: Sclera segmentation benchmarking competition in the mobile environment. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE.

Vitek, M., Hafner, A., Peer, P., and Jaklič, A. (2021). Evaluation of deep approaches to sclera segmentation. In *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1097–1102.

Wang, C., He, Y., Liu, Y., He, Z., He, R., and Sun, Z. (2019). Sclerasegnet: an improved u-net model with attention for accurate sclera segmentation. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE.

Wang, C., Muhammad, J., Wang, Y., He, Z., and Sun, Z. (2020a). Towards complete and accurate iris segmentation using deep multi-task attention network for non-cooperative iris recognition. *IEEE Transactions on Information Forensics and Security*, 15:2944–2959.

Wang, C., Wang, Y., Liu, Y., He, Z., He, R., and Sun, Z. (2020b). Sclerasegnet: An attention assisted u-net model for accurate sclera segmentation. *IEEE transactions on biometrics, behavior, and identity science*, 2(1):40–54.

Wang, Z., Feng, Y., and Tao, Q. (2010). Momentum based level set segmentation for complex phase change thermography sequence. In *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, volume 12, pages V12–257–V12–260.

Wilcoxon, F., Katti, S., and Wilcox, R. A. (1970). Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259.

Wildes, R. P. (1997). Iris recognition: an emerging biometric technology. *Proceedings of the IEEE*, 85(9):1348–1363.

Wolf, L. and Bileschi, S. (2006). A critical view of context. *International Journal of Computer Vision*, 69(2):251–261.

Xiao, L., Sun, Z., and Tan, T. (2012). Fusion of iris and periocular biometrics for cross-sensor identification. In *Biometric Recognition*, pages 202–209.

Zanlorensi, L. A., Laroca, R., Lucio, D. R., Santos, L. R., Britto Jr., A. S., and Menotti, D. (2020). UFPR-Periocular: A periocular dataset collected by mobile devices in unconstrained scenarios. *arXiv preprint*, arXiv:2011.12427:1–12.

Zanlorensi, L. A., Laroca, R., Luz, E., Britto Jr., A. S., Oliveira, L. S., and Menotti, D. (2019). Ocular recognition databases and competitions: A survey. *arXiv preprint*, arXiv:1911.09646:1–20.

Zanlorensi, L. A., Lucio, D. R., Junior, A. d. S. B., Proença, H., and Menotti, D. (2020). Deep representations for cross-spectral ocular biometrics. *IET Biom.*, 9(2):68–77.

Zanlorensi, L. A., Luz, E., Laroca, R., Britto, A. S., Oliveira, L. S., and Menotti, D. (2018). The impact of preprocessing on deep representations for iris recognition on unconstrained environments. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 289–296. IEEE.

Zanlorensi, L. A., Proença, H., and Menotti, D. (2020). Unconstrained periocular recognition: Using generative deep learning frameworks for attribute normalization. *arXiv preprint*, arXiv:2002.03985:1–5.

Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., and Agrawal, A. (2018). Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7151–7160.

Zhang, H., Xue, J., and Dana, K. (2017). Deep ten: Texture encoding network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 708–717.

Zhang, J., Xie, C., Xu, X., Shi, Z., and Pan, B. (2020). A contextual bidirectional enhancement method for remote sensing image object detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:4518–4531.

Zhang, W. and Ma, Y. (2014). A new approach for iris localization based on an improved level set method. In *2014 11th International Computer Conference on Wavelet Actiev Media Technology and Information Processing(ICCWAMTIP)*, pages 309–312.

Zhang, Y., Yuan, Y., Feng, Y., and Lu, X. (2019). Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5535–5548.

Zhao, Z. and Kumar, A. (2019). A deep learning based unified framework to detect, segment and recognize irises using spatially corresponding features. *Pattern Recognition*, 93:546–557.

Zhou, L., Ma, Y., Lian, J., and Wang, Z. (2013). A new effective algorithm for iris location. In *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1790–1795.

Zhou, Z., Du, E. Y., Thomas, N. L., and Delp, E. J. (2012). A new human identification method: Sclera recognition. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 42(3):571–583.

Zhu, Y., Tan, T., and Wang, Y. (2000). Biometric personal identification based on iris patterns. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, pages 801–804 vol.2.

Zuo, J., Schmid, N. A., and Chen, X. (2007). On generation and analysis of synthetic iris images. *IEEE Transactions on Information Forensics and Security*, 2(1):77–90.