Cartography M.Sc.

Master thesis

Machine Learning Image Segmentation to Improve Object Recognition in Mixed Reality

Guillermo Fernando Esquivel Tabares

Technical University of Munich

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

TECHNISCHE UNIVERSITÄT DRESDEN

ITC UNIVERSITY OF TWENTE.

2020

Machine Learning Image Segmentation to Improve Object Recognition in Mixed Reality

submitted for the academic degree of Master of Science (M.Sc.)
conducted at the Department of Aerospace and Geodesy
Technical University of Munich

Author: Guillermo Fernando Esquivel Tabares
Study course: Cartography M.Sc.
Supervisor: Dr.-Ing. Christian Murphy (TUM)
Reviewer: Dr. Rania Kounadi (UT)

Chair of the Thesis
Assessment Board: Prof. Dr. Liqiu Meng

Date of submission: 11.10.2020

Statement of Authorship

Herewith I declare that I am the sole author of the submitted Master's thesis entitled:

"Machine Learning Image Segmentation to Improve Object Recognition in Mixed Reality"

I have fully referenced the ideas and work of others, whether published or unpublished. Literal or analogous citations are clearly marked as such.

Munich, 11.10.2020 Guillermo Fernando Esquivel Tabares

Acknowledgments

First of all, I would like to express a sincere appreciation to the people who contributed to building this research. To Christian Murphy for his constant feedback, to Juliane Cron for her ideas from a cartographic perspective, to Rania Kounadi for her initial remarks to improve the development of my investigation, and to Ross Clark for his detailed review.

Likewise, I want to thank Johan Avendaño as well as Miguel Arjona and their students for participating in the user tests.

Special recognition is also due to the people who anonymously comment and answer questions in technology forums. I could hardly have solved all the obstacles I faced without their selfless and generous help.

Finally, these are difficult times all over the world, however, it all makes sense when you have people to love. So, I want to use this space to thank life for my friends in Mexico, Colombia, Germany, and Sri Lanka, for my family, and for Miguel, for always giving me reasons.

Abstract

Object recognition represents an emerging technology in the field of image processing able to detect and label objects through the recognition of patterns in images. At the same time, Mixed Reality represents the combination of the virtual and physical worlds in a bid to yield a digital environment where elements from both dimensions co-exist. Through the integration of an image segmentation algorithm along with image enhancement techniques, this thesis aims to facilitate the navigation experience in Mixed Reality by recognizing more efficiently those objects that provide relevant information to users to navigate. The image segmentation algorithm and the image enhancement techniques are implemented in a video recording, in such a way that through the detection and modification of object features, their instances are either visually highlighted or downgraded according to the information they provide to fulfill the navigation task. Subsequently, in order to determine the impact on human perception, two user tests are conducted. In the first test, users are asked to focus their attention on a virtual element and select the objects that attract their attention the most. In the second test, in which the methodology of this thesis is implemented, users are also asked to focus on a virtual element added to the video and choose the elements that are most striking to them. The results show that the technique used to highlight objects allowed users to recognize them more easily. In contrast, the objects that were downgraded remained eye-catching to users.

Keywords: Navigation; image segmentation; object recognition; Machine Learning; Mixed Reality; Visual enhancement.

Kurzfassung

Die Objekterkennung stellt eine aufkommende Technologie im Bereich der Bildverarbeitung dar, die in der Lage ist, Objekte durch die Erkennung von Mustern in Bildern zu detektieren und zu beschriften. Gleichzeitig stellt Mixed Reality die Kombination der virtuellen und der physischen Welt dar, um eine digitale Umgebung zu schaffen, in der Elemente aus beiden Dimensionen nebeneinander existieren. Durch die Integration eines Bildsegmentierungsalgorithmus zusammen mit Bildverbesserungstechniken zielt diese Arbeit darauf ab, die Navigationserfahrung in der gemischten Realität zu erleichtern, indem diejenigen Objekte effizienter erkannt werden, die den Benutzern relevante Informationen zur Navigation bieten. Der Bildsegmentierungsalgorithmus und die Bildverbesserungstechniken werden in einer Videoaufzeichnung so implementiert, dass durch die Erkennung und Modifikation von Objektmerkmalen deren Instanzen entweder visuell hervorgehoben oder entsprechend der Informationen, die sie zur Erfüllung der Navigationsaufgabe liefern, herabgestuft werden. Anschließend werden zwei Benutzertests durchgeführt, um die Auswirkungen auf die menschliche Wahrnehmung zu ermitteln. Im ersten Test werden die Benutzer gebeten, ihre Aufmerksamkeit auf ein virtuelles Element zu richten und die Objekte auszuwählen, die ihre Aufmerksamkeit am meisten auf sich ziehen. Im zweiten Test, in dem die Methodologie dieser Arbeit umgesetzt wird, werden die Benutzer ebenfalls gebeten, sich auf ein virtuelles Element zu konzentrieren, das dem Video hinzugefügt wurde, und die Elemente auszuwählen, die ihnen am meisten auffallen. Die Ergebnisse zeigen, dass die zur Hervorhebung von Objekten verwendete Technik es den Benutzern ermöglichte, diese leichter zu erkennen. Im Gegensatz dazu blieben die Objekte, die herabgestuft wurden, für die Benutzer auffällig, jedoch in geringerem Maße.

Schlüsselwörter: Navigation; Bildsegmentierung; Objekterkennung; Maschinelles Lernen; Mixed Reality; Visuelle Verbesserung.

Table of contents

List of figures

List of tables

1. Introduction

1.1. Motivation and problem statement

According to Fallah (2010, p. 24), "The ability to navigate effectively and safely in unfamiliar environments relies upon being able to build a cognitive map of the environment." Rokhsaritalemi et al. (2020) recognize the importance of constructing optimal visualizations when navigating to contribute to making appropriate decisions. Upholding these statements, Lorenz et al. (2015) assert that it is vital that users have the feeling of being involved in a real ambiance as using built models, encouraging the labor of getting a complete understanding of the surroundings to assist navigation in an effective way (Giesemann et al., 2017).

In this sense, Mixed Reality (MR) represents an useful mechanism that provides users with a more detailed context of their environment (Grasset et al., 2011). *MR* is thought of as the combination of the virtual and the real in a bid to yield a digital world where elements from both dimensions co-exist. Although there is no universal consensus over these terms, *MR* spans Augmented Reality (*AR*) and Virtual Reality (*VR*), being considered advanced techniques to help users navigate through digital models (Tadros and Franklin, 2019). Further approximations, such as Çöltekin et al. (2020), consider that *MR* includes issues of *AR* and vice-versa, and elaborate the notion of Expanded Reality, which comprises *MR*, *AR*, and *VR* as immersive technologies to develop visualizations. According to Nagata et al. (2017), the combination of virtual and real representations in one single interface represents the foremost advantage provided by this technology.

In this way, this thesis aims to establish a methodology that makes it easier to detect certain objects in a mixed reality context, so that the task of navigating is performed more efficiently for the user. This process consists of an image segmentation model capable of detecting and extracting the instances of elements displayed during a video recording, and image enhancement techniques that highlight and downgrade their visual properties.

For its part, image segmentation embodies the extraction of meaningful regions from imagery by putting into effect a pixel sorting task based on its features and the characteristics of its neighbors (Iglovikov and Shvets, 2018). In line with Haralick and Shapiro (1985), image segmentation can be understood as a clustering process of linkable, growing regions created from the pixel gray intensity and their distinctiveness from the background.

Feature extraction and object detection models, supported by *Machine Learning* algorithms, are essential to fulfill image segmentation by distinguishing elements, offering distinctiveness to each part being perceived, and training as well as labeling data in accordance with the probability of belonging to a known class (Hore et al., 2018). This way, segmentation and classification in images have spurred scene understanding and supplied thorough information concerning location, shape, and number of individual objects (Hayder et al., 2017).

Regarding image enhancement techniques, they aim to make certain parts of the image more noticeable to the user by modifying the properties of the pixels (Kaur, 2013; Maini and Aggarwal, 2010). Similarly, the opposite process also takes place from the distortion of both pixel features and image regions, making it more difficult to recognize certain objects and redirecting the viewer's attention to clearer regions of the image (Murphy, 2015).

Consequently, the combination of both image segmentation models and image enhancement techniques in *MR* could lead to an improvement of the navigation experience by making it easier to grasp the information available in a certain context. Likewise, this approach addressing both image processing and visual design techniques has not been profoundly explored, bringing the opportunity to establish a state-of-the-art methodology for object recognition in the navigation field.

1.2. Research identification

The current thesis aims to facilitate the navigation experience in Mixed Reality through the detection and extraction of instances of objects, which will be highlighted or downgraded according to the level of information they provide. Forthcoming developments of this approach could lead to building enhanced real-time navigation tools.

1.2.1. Research objectives

1.1.1.1. Calibrate and parse an adequate object recognition model able to generate instance segmentation in a video.

1.1.1.2. Implement image enhancement techniques to highlight/downgrade elements on the scene.

1.1.1.3. Integrate and test the image segmentation output with the image enhancement techniques.

1.2.2. Research questions

- What is the difference between instance segmentation and object detection?
- Which existing model fits adequately to generate the desired segmentation process?
- How to integrate an object recognition model with image enhancement techniques to highlight and downgrade elements in a video?
- Can a preprocessed image segmentation tool together with visual enhancement methods facilitate object recognition in *Mixed Reality?*

1.3. Innovation aimed at

This thesis aims at introducing a new approach to tackle navigation by addressing components of object recognition and image enhancement techniques. Image processing and visual enhancement models will be used combined to show how they can strengthen navigation through real scenarios. Although the scope of this thesis will be applied to already recorded moving scenes, in forthcoming developments it could be brought to real-time execution and integrated with existing navigation tools.

2. Conceptual framework

2.1. Machine learning and deep learning

As stated by Iglovikov and Shvets (2018), the most straightforward method to detect and classify objects in imagery is manually; however, this method is time-consuming and might suffer inconsistencies due to human inaccuracies. Therefore, as the same authors argue, it is necessary to automatize the treatment of images as they are produced, so that machines accomplish the job. Nowadays, advances in hardware components such as accelerated GPUs (graphics processing unit) allow executing complex algorithms that learn from massive image datasets with efficient extraction and generalization capabilities and, consequently, can conduct very precise object classification (Zhu et al., 2020).

In this sense, Artificial Intelligence (*AI*) provides to machines the ability to learn from their environment in order to fulfill a determined request. Franca et al. (2019) highlight that Machine Learning (*ML*) constitutes a branch of *AI* with the capacity, through mathematical models and pattern recognition, to make decisions and perform activities with no human intervention. Therefore, *ML* works by learning from input and output data, which may or may not be labeled, as follows:

1) *Active learning* overcomes data labeling by searching unlabeled instances that can be sorted by an observer (human or machine), minimizing the effort of providing tagged inputs (Souza et al., 2017).
2) *Supervised learning* is based on the relationship between a given input and output dataset. In such a case, data is entirely labeled (Liu and Wu, 2012).
3) *Unsupervised learning* strives to extract valuable information from huge amounts of unlabeled data, by learning from significant features (Chen et al., 2016).

In addition, Deep Learning (*DL*) responds to a derivation of machine learning that allows computational frameworks to understand and learn from complex data, showing outstanding results in object recognition, language processing, and medical image examination (Liu et al., 2020). Likewise, Zhao et al. (2019) acknowledge that the leading contribution of *DL* comes from training algorithms with large datasets containing up to millions of inputs, accelerating and improving the image, video, and voice recognition processes.

2.1.1. Convolutional Neural Network structure

Zhao et al. (2019) assert that object recognition deals with two main points when identifying elements: 1) where they are located in a given image subset and 2) which category/class they belong to. To carry out these assignments, Convolutional Neural Networks (*CNN*) take place as an automatic system capable of learning from massive datasets and solving pattern recognition difficulties (O'Shea and Nash, 2015).

On one hand, Liu et al. (2020) illustrate *CNN* as the *DL's* most important model of learning, composed of several layers extracting information from data with different levels of abstraction. Similarly, O'Shea and Nash (2015) affirm that *CNN* represents a form of Artificial Neural Network (*ANN*), a scheme that works as a biological brain at the moment of processing information (Zhu et al., 2020), aiming to learn from inputs and boosting a final required outcome.

Although there exist different approximations to deal with *CNN*, most agree with the fundamental steps needed to go through the whole *CNN* methodology and get the expected result. Figure 1 illustrates the overall components of a common *CNN* structure based on several authors' approaches (Albawi et al., 2017; Liu et al., 2020; O'Shea and Nash, 2015; Zhu et al., 2020; Venugopal et al., 2019). This arrangement shows the path of an input image through a convolution process in which each part of it is evaluated to identify the class the object has the greatest probability to be part of.

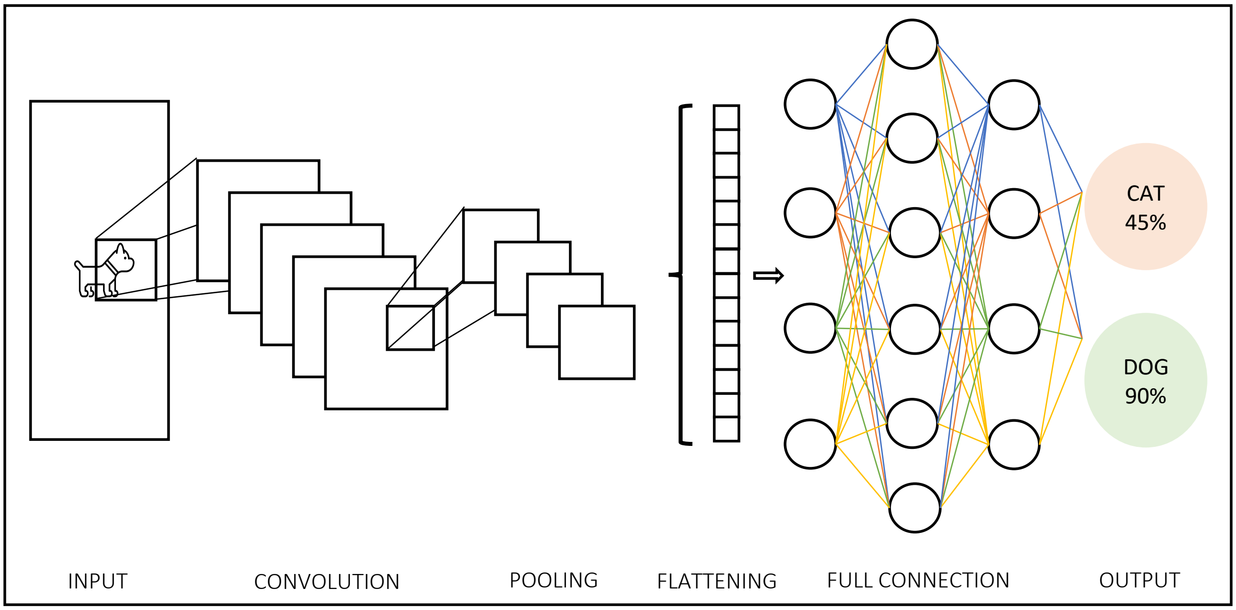


Figure 1. Structure of a simple Convolutional Neural Network.
Source: Author design based on O'Shea and Nash (2015) and Zhu et al. (2020).

A simple *CNN* works through the use of a set of arbitrary size filters called feature maps, which are responsible for extracting attributes from neighboring pixels by scanning segment by segment through the whole image (Zhu et al., 2020). As Figure 1 demonstrates, feature maps are applied many times as convolutional layers, involving responses over local areas and identifying their overall spatial location (Zhao et al., 2019). Once all segments are organized in a vector stance, as illustrated in the flattening step, each neuron (circles) of the full connection estimates a score in relation to the class in which the object falls into, returning the label, spatial position, and extension of the corresponding category (O'Shea and Nash, 2015).

2.1.1.1. Convolution

As indicated by O'Shea and Nash (2015), every image is taken as an input layer composed of pixels holding radiometric values distributed, usually, in three channels: red, green, and blue. The convolution layer performs the first activities of the *CNN*, under which each band of the input layer is convolved by sliding windows, previously mentioned as feature maps. The convolution task is executed as many times as required, searching for and pulling out certain features, and then stacking the outcome sublayers as the exploration through the image is expanded (Zhu et al., 2020).

2.1.1.2. Pooling

By pooling, the complexity for further layers through the CNN is reduced. This labor consists of down-sampling of the feature maps without modifying the number of filters (Liu et al., 2020). Albawi et al. (2017) see pooling as an image resolution diminution, Max-pooling being the most common type, working, commonly, with a 2x2 mask arrangement. The mask breaks the incoming sublayer into subregion blocks, taking the maximum pixel value from it, as Figure 2 exemplifies.

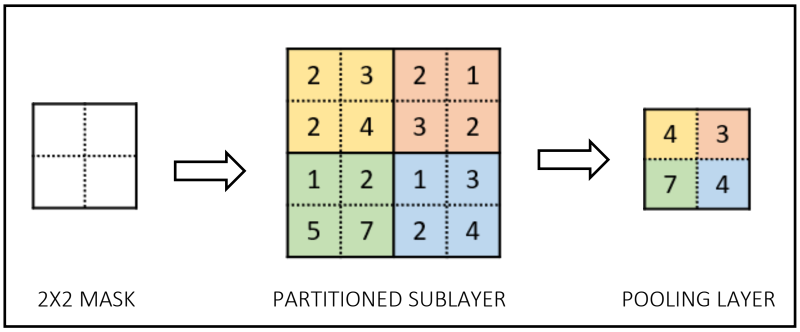


Figure 2. Pooling execution using a 2x2 mask.
Source: Author design based on Albawi et al. (2017).

2.1.1.3. Flattening

Once the convolution and pooling are completed, the generated values, organized in arrays, are transformed into a vector shape, as exposed in Figure 3. According to Venugopal et al. (2019), the arrays contain nongraphical inputs and features from the images that are appended to this vector which is thereupon processed via the following full connected layer. The general idea of flattening is to facilitate the understanding of the image in further convolutional layers by reducing the original number of dimensions to one.

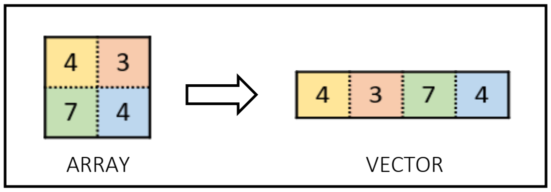


Figure 3. Flattening process.
Source: Author design based on Venugopal et al. (2019).

2.1.1.4. Full connection

Fully connected layers perform the last stage towards object recognition. In this event, the structure is integrated by totally connected nodes in both directions, back and forward, which are trained to identify specific characteristics from the incoming vector layer (Albawi et al., 2017). In such circumstances, each node evaluates the input and assigns a score that is used for classification, predicting the object label and its location simultaneously (O'Shea and Nash, 2015; Zhao et al., 2019). As a result, the corresponding output will contain information regarding the label of the target object, its features, and spatial information (Zhu et al., 2020).

2.2. Image segmentation

Image segmentation is the process of yielding clustered subregions to identify sections of interest in imagery, which are used in further image processing stages to simplify the analysis and understanding of the whole scene (Zhu et al., 2020). In other words, segmentation refers to the splitting of a digital image into a set of pixels with common visual features, in order to define the boundaries of the objects displayed (Singh and Singh, 2010).

Similarly, Swaminathan et al. (2020) define image segmentation as a region-based tracing method that serves as a pre-process step towards complex computer vision tasks, including object recognition. Supporting this stance, Zhu et al. (2020) concede that the foremost intention of segmenting images is to dissociate a target object data from the original image, so as to ease the object identification process.

Some authors, such as Haralick and Shapiro (1985), point out that an appropriate image segmentation process should care about maintaining significant value differences between each region, bringing simplicity and providing spatial accuracy to the unit drawn. In this sense, Singh and Singh (2010) underline that the result of performing image segmentation could mean either a set of pieces covering the whole picture or a group of contours drawn on selected features.

2.2.1. Object recognition

Object recognition is defined as a set of subtasks that provide a semantic understanding of digital images and footages aiming to ascertain the identity of elements based on known predefined labels (Yang, 2009; Z.-Q. Zhao et al., 2019). Similarly, Liu et al. (2020) describe object recognition as the determination whether an object exists in a digital visualization, so that the instance of its shape and spatial location can be obtained.

Some approaches, such as that of Gould et al. (2009), consider that image segmentation and object recognition are acutely related actions, indeed combined by integrating the pixel, region, and object analysis in the same methodology. This approximation determines two groups of the scene: background and foreground. Those pixels belonging to the background are thought of as part of the landscape (sky, buildings, relief, etc.), and those in the foreground correspond to the object itself.

According to Hariharan et al. (2014), object recognition can be divided into two main categories based on their outcomes: object detection and semantic segmentation. Object detection produces bounding boxes around identified objects. In contrast, semantic segmentation marks out only those pixels that are part of the object, getting exclusively the geometry that covers it.

2.2.1.1. Object detection

As stated by Hariharan et al. (2014), object detection stands for an approach of object recognition. This approximation aims at localizing individual objects in digital images to subsequently demarcate them into fitting bounding boxes along with labels from a set of predefined categories (such as people, table, dog, etc.) (Miksys et al., 2019). Furthermore, Liu et al. (2020) indicate that returning objects enclosed into geometries, pinpoints the extension as well as the spatial location of elements present in the image, providing more context from the scene.

Figure 4 shows the outcome of yielding object detection. In the example, the contours and labels display for each object both its spatial extension and the category it belongs to. In accordance with Franca et al. (2019), this method works through the search for recognizable patterns that can be associated with a class of objects that are known by a recognition model.

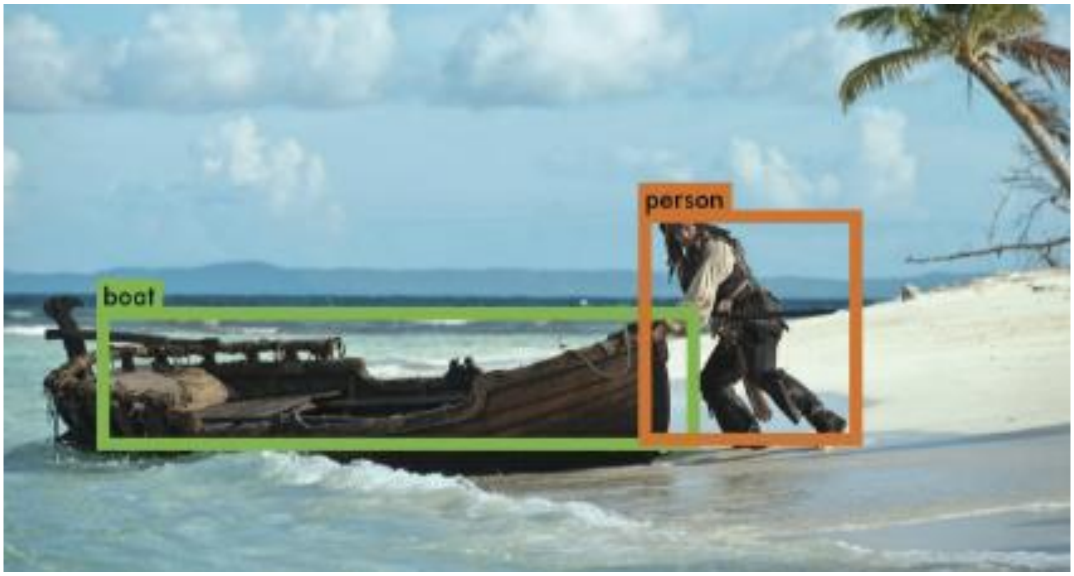


Figure 4. Example of object detection approach.
Source: Redmon et al. (2016).

2.2.1.2. Semantic or instance segmentation

Unlike object detection, instance segmentation grinds out a geometric shape or mold containing exclusively those features that the object inheres within it. This method supplies a detailed understanding of the scene by detecting individually all objects, segmenting each instance of them, and providing their location and shape (Hayder et al., 2017; Z.-Q. Zhao et al., 2019). Miksys et al. (2019) set forth that this approach stands out since it offers a map at the pixel level of the element being detected.

Figure 5 shows an example of instance segmentation. As can be seen, the result of this method is a group of regions that cover the shapes of the objects individually. The example also shows a set of boxes where the objects are contained, however, this feature is not part of this model.

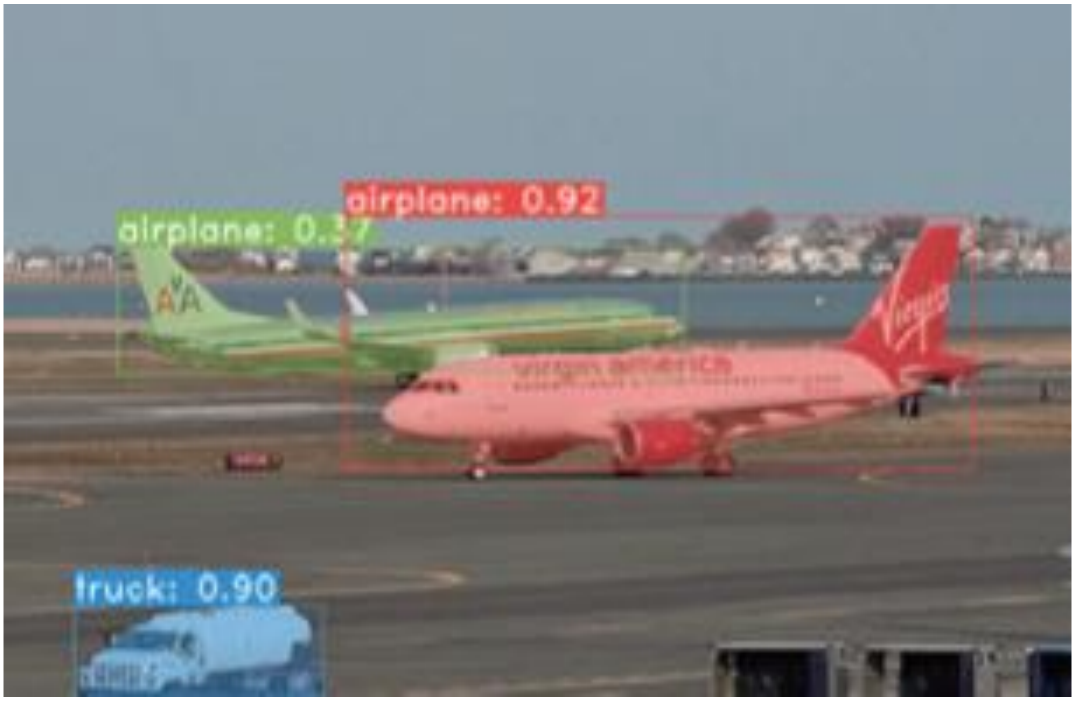


Figure 5. Example of instance segmentation approach.
Source: Bolya et al. (2019).

2.2.2. Related work

Computer vision is mastering a huge set of image processing domains to get the most out of visual representations. One of them corresponds to object recognition which has made much progress since the use of *CNN* (Hayder et al., 2017). Currently, object recognition is adapted to diverse areas such as drone navigation, autonomous driving, robotic manipulation, and plant analytics (Miksys et al., 2019).

Usual object recognition systems apply a two-step computing process. Firstly, a reckoning over the image is done to extract contained features and, secondly, a classification runs to identify what elements are present. Typically, these techniques employ Machine Learning and Deep Learning applications for setting a classifier that learns from a training dataset fed with images (Giesemann et al., 2017).

In the case of navigation, many algorithms are used to assist this task by extracting information from the scene. *Scene labeling* is one of the examples; it identifies objects by assigning each pixel from an input image to a label corresponding to an object (Giesemann et al., 2017). Object recognition under this method is made by encompassing the relationships among objects and enabling obstacle detection, close-range estimation, and relative positioning location.

The development of models used for object recognition has accelerated and brought accuracy to the delimitation of objects, so they can be easily differentiated from non-relevant information (Liu et al., 2020). For instance, Hariharan et al. (2014) introduced an additional step to *CNN* called *Region refinement*, allowing the suppression those additional pixels belonging to different classes in semantic segmentation, thus they could calculate a more precise individual mask of the element to be detected.

In the chain of progress, *YOLO* (You Only Look Once), presented in 2016, was created to perform object detection in real-time by mining features, predicting bounding boxes, and assigning scores to categories (He, 2016; Redmon et al., 2016). Further developments of *YOLO* are *YOLOv2* (Nakahara et al., 2018), *YOLO9000* (Redmon and Farhadi, 2017), *YOLOv3* (Redmon and Farhadi, 2018), and *YOLOv4* (Bochkovskiy et al., 2020).

Regarding semantic segmentation, an example is the Pyramid Scene Parsing Network, which is a semantic-based segmentation model that assigns each pixel in the image to a category of objects (Zhao et al., 2017). Thus, this model provides the shape of each object, the label of the class it belongs to, and its location in the image. In the same way, *YOLACT* (You Only Look at Coefficients), which was brought to light in late 2019, is an object recognition model capable of producing semantic segmentation and object detection simultaneously, yielding high-quality object masks and bounding box containers for each object (Bolya et al., 2019).

In summary, object recognition represents an exploratory field for scene understanding and data mining. The alternatives to identify elements in the scene go through a variety of classifiers able to return the category where the item has the greatest probability to belong to. Object detection and instance segmentation represent the approaches in which object recognition is generated, either by returning a bounding box enclosing the object or by marking off its very boundaries.

2.3. Mixed Reality

As stated by Speicher et al. (2019), nowadays many definitions regarding *MR* have been delivered to the academic world, contributing to confusion as well as imposing concepts. In this document, based on sundry approaches, *MR* will be treated as a set of technologies that allow combining elements from both the actual and digital worlds, also incorporating methods to alter image depiction, such as blurring together with all those changes of image properties carried out by computer-assisted techniques (Costanza et al., 2009; Rokhsaritalemi et al., 2020; Speicher et al., 2019; Tadros and Franklin, 2019).

Notwithstanding, some approximations will be mentioned to construct a holistic view of this term. As stated in Rokhsaritalemi et al. (2020), *MR* is considered a merger of real and virtual dimensions, mixing elements from both to produce practical scenarios to users. Accordingly, Costanza et al. (2009) interpret *MR* as the superposition of computer-generated objects onto images from the real world, giving users the perception of their physical context along with virtual elements that might also be appreciated as visually modified parts of the scene (using semitransparency, for instance). Rokhsaritalemi et al. (2020) list the following aspects as the main features conforming any *MR* system:

1) The combination of objects from the real and virtual world.
2) The interaction with users in real-time.
3) The connection between virtual and real elements.

Further approaches, such as Çöltekin et al. (2020), think of *MR* as a hybrid concept that encompasses everything that falls into reality and virtuality, with *VR* and *AR* being its most prominent branches. Likewise, Kunkel and Soechtig (2017) contemplate *MR* as an encounter between *VR* and *AR*, bringing together a dimension where data from digital and physical objects coexist and interact. Additionally, in recent studies such as Çöltekin et al. (2020), Fast-Berglund et al. (2018), and Mann et al. (2018), the term of Extended Reality (*ER*) came up to cover all combinations between artificial and real contexts, spanning *MR*, *AR*, and *VR*, as Figure 6 details.

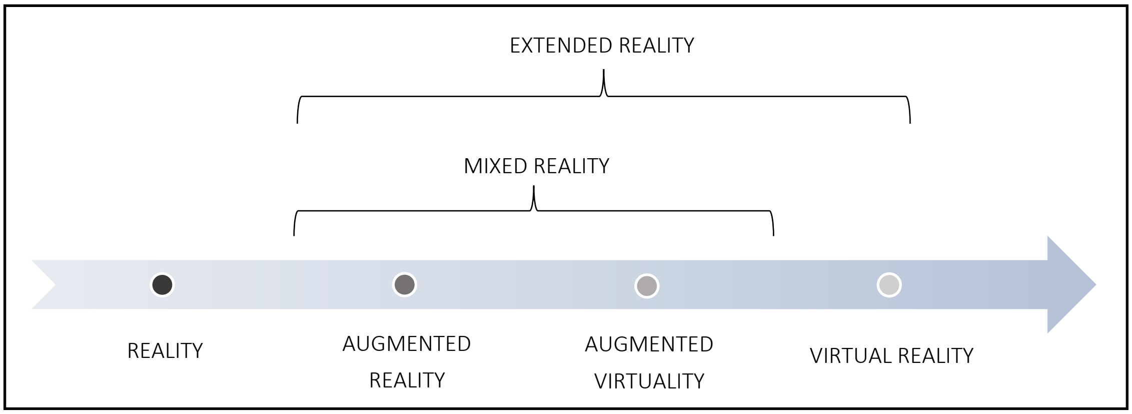


Figure 6. Dimensions covered by Mixed Reality and Extended Reality.
Source: Author design based on Çöltekin et al. (2020), Fast-Berglund et al. (2018) and Mann et al. (2018).

2.3.1. Extended Reality

As mentioned in the introduction, the relationship between *MR*, *VR* and, *AR*, has not achieved universal consensus. Therefore, more concepts are brought to light to help understand the association between those terms. Fast-Berglund et al. (2018) explore the notion of *ER*, which covers all mixed virtual-real environments and human-machine interrelations yielded by computers. In this situation, *AR* and *VR* are seen as extensions of *MR* that fall in between a physical space representation and a cyber system. Fast-Berglund et al. (2018) also include the category of Augmented Virtuality which, in concordance with *AR*, performs a higher level of digitalization by displaying even more synthetic objects.

Mann et al. (2018) use the shorthand *xR* as well to name Extended Reality. *xR* stands for all technologies that extend human capabilities to understand the physical world by employing computational methods. Likewise, Çöltekin et al. (2020) utilize the approach of Immersive Technologies (also referred to as *xR*) to call those techniques spanning analytic visualizations, modified reality, and human-computer relationship. Figure 6 deploys the spectrum that leads from reality to virtuality in *xR*, taking into account that "x" symbolizes the axis overcoming reality, interpolating between the material and virtual nature.

2.3.2. Virtual Reality

As previously stated, there is still no agreement regarding the derivation and relationship between *MR* and *VR*. Costanza et al. (2009) relate the conceptual and historical birth of *MR* to the development of *VR* in the '60s. Different points of view, such as Çöltekin et al. (2020) claim that *VR* along with *AR*, and all those combinations of tangible and digital environments are direct subcategories of *MR*. Even though no unique definition has settled, the current thesis will consider *VR* as well as *AR* as *MR* branches. Costanza et al. (2009) understand *VR* as a set of computing tools that generate utterly virtual environments wherein the user is immersed through not only their visual capabilities but also audible and tangible.

The digital surroundings made by *VR* provide users with the feeling of being isolated and ringed by a real context in a computer-based setting. In the same way, Çöltekin et al. (2020) express that *VR* aims to recreate scenarios where humans would not be able to distinguish from real to artificial experiences. Since an adequate *VR* asset should stimulate all senses, today's advances in audio and visualization have made huge progress, but touching, tasting, and smelling continue to be sensations in which lots of work needs to be done.

2.3.3. Augmented Reality

Azuma (1997) defines *AR* as a variation of *VR* that stands for the integration and exhibition of three-dimensional virtual objects into a real environment, enabling them to be displayed in real-time. Çöltekin et al. (2020) put this concept in simple words by saying that *AR* is the superposing of virtual objects onto views from the actual world.

Rokhsaritalemi et al. (2020) affirm that *VR* requires users to use additional devices to interact with an absolute digital medium, which might represent a drawback since there is a lack of connection with the actual world. On the other hand, *AR* is seen as a visual compensation for that missing link, by integrating computer-generated features with real content. Mann et al. (2018) point out that *MR* is an unintentionally modified reality where the actual world is not intended to be changed but overlaid by cyber objects.

As shown in Figure 6, *AR* represents the first stage from reality to a complete digital recreation by allowing virtual elements to be overlapped onto views from the real world (Çöltekin et al., 2020). Mann et al. (2018) also explain *AR* as being similar to *VR* but instead of overlooking reality, digital-based content is added to the real experience. As displayed in Figure 7, *AR* is mostly composed of the perceived world along with some artificial elements, providing useful information to help navigate (Knutsson and Georgsson, 2019).

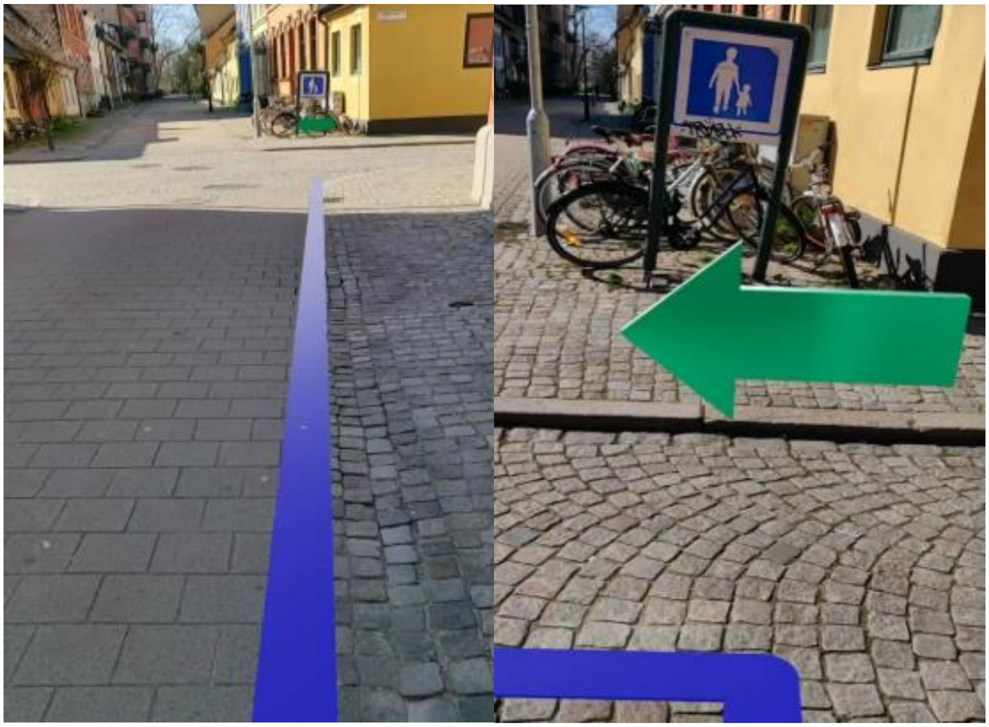


Figure 7. Augmented Reality in human navigation.
Source: Knutsson and Georgsson (2019).

2.3.4. Related work

At the same time, many tools have been created to improve navigation through *MR*, *VR*, and *AR* mechanisms. These technologies have been under development for more than twenty-five years (Nagata et al., 2017). Tadros et al. (2019) state that *MR* has had and will have a significant impact in many industry fields by allowing users to navigate facilities and environments built upon 3D models and holograms. Besides navigation, *MR* has contributed to areas such as advertising, entertainment, education, medical purposes, and mobile apps (Carmigniani et al., 2011).

To illustrate the application of this system, Knutsson and Georgsson (2019) assembled an Android app to guide pedestrians from two points by combining *AR* with a phone's integrated GPS*.* The framework used to build the prototype was based on *Google's ARCore*, a platform for setting up *AR* applications, and *Sceneform*, a library for managing 3D objects. These two schemes are exclusively available for android applications though, bringing up some constraints for their use.

An interesting element of this research was the confrontation between two- and three-dimensional tools for human navigation in mobile devices. Participants in this study agreed that instructions from a 3D representation were easier to follow than those in a 2D view since the deployed abstraction required a minor effort to be interpreted and, simultaneously, there was no space for confusion. Regarding disadvantages of using *AR*, users stated that it was hard to plan the overall route, they were more screen dependent and manual calibration was necessary to compensate for sensors' miscalculations (Knutsson and Georgsson, 2019).

Likewise, Mehdi et al. (2020) address the use of open Python-based libraries such as *OpenCV*, *OpenGL*, and *OpenCL* to render forthcoming edifices to be built on a real city landscape, by overlapping 3D figures on a real-time video stream. Nowadays, high-resolution videos up to 4*K* support *AR* operations with the aforementioned image processing libraries, as shown in Shin et al. (2020).

Furthermore, a wide number of studies in various scientific fields have taken the benefits from these virtual environments, and not always involving a visual representation. Fallah (2010) shows the implementation of *MR* on a system to help the visually impaired navigate indoors by sending signals to the user's cellphone when an obstacle was ahead. Although no visualization associated method is reached in this investigation, it relies upon three-dimensional models in a virtual world.

The progress made in *MR* has modified the manner in which people interact with technology by easing decision making. In this way, the usefulness of *MR* is getting more attention. A recent Goldman Sachs study declared that by 2025 *MR* will become an $80 billion industry, overcoming gaming and entertainment (Medici, 2016). Further developments of *MR* will affect all everyday life areas, and geospatial sciences are incorporating *AR* and *VR* as immersive technologies helping users get a more understandable context of their environment.

2.4. Image enhancement

As one of the areas belonging to image processing, visual enhancement oversees the process of making certain features of interest in an image more obvious to the observer (Kaur, 2013). Usually, image perception improvement is carried out by modifying some pixel properties such as contrast, brightness, saturation, and so on. According to Maini and Aggarwal (2010), enhancement techniques can be assigned to two main categories: 1) Spatial domain methods, which manipulate pixel values and 2) frequency-domain methods, where the image is treated in its frequency domain.

From Kaur (2013) and Maini and Aggarwal (2010) approximations, based on the human visual mechanism as well as observer's experiences and perspectives, visual enhancement in images responds to a subjective field. In this sense, the effectiveness of image enhancement processes needs to meet the expectations of the user in order to fulfill a determined assignment.

2.4.1. Highlighting methods

Highlighting can be understood as the alteration of image characteristics to visually emphasize objects contained in it (Murphy, 2015). Accordingly, Maini and Aggarwal (2010) recognize this process as a transformation whereby the pixels of an input image are put through a mathematical expression that changes their original attributes. Pursuant to Kaur (2013), Saleem et al. (2012), and Yoon et al. (2009), contrast, image brightness, and saturation, correspond with the main operations to perform pixel intensity variation in conformity with the field in which the image enhancement is required.

2.4.1.1. Contrast

Iwasokun and Akinyokun (2014) define contrast as the gray value difference between two neighboring pixels. Similarly, Kaur (2013) points out that the contrast of an image is given by the range between its highest and lowest intensity value. The idea behind this approach is to get a clearer image to eyes through an intensity value redistribution, making objects in the scene easier to distinguish (Yoon et al., 2009). Additionally, Saleem et al. (2012) state that contrast enhancement is a mechanism that allows improving visibility of details without adding unpleasant artifacts in the picture.

2.4.1.2. Brightness

Although the concepts of contrast and brightness might be mixed up, they do not represent the same characteristic. An image depicting a pale object with a white background has high brightness but low contrast and, contrariwise, the same object with a dark background represents low brightness with high contrast. Nimkar et al., (2013) mentions that the change of contrast affects the portrayal of dark and bright regions by making them brighter or darker. Therefore, unlike contrast, brightness deals with the overall lightness of the picture, *Gamma* being the unit of measure (Smith, 1999).

2.4.1.3. Saturation

Saturation is seen as a measure of color intensity (Van Hurkman, 2011). As mentioned by Nimkar et al. (2013), saturation yields a clear definition between shadows and highlights. Zhang (2017) illustrates the process of saturation enhancement by arguing that any amplification involving a factor greater than one can make the color more notorious (brighter) or, in the opposite way, by multiplying by a number smaller than one the sensitivity of the color is diminished.

2.4.2. Downgrading

Another approach to highlight portions of the image is made by reducing the role of those elements that might attract the user's attention. This operation allows redirecting concentration to parts of the image where the user needs to be focused on. In this case, the downgrading method is applied over the object to be debased, so their attributes are harder to identify and perceived as part of the background (Murphy, 2015).

2.4.2.1. Blurriness

Sandford et al. (2018) describe blurring as the method to take the image to its lowest spatial frequency, generating a degradation of details perception whereby the recognition task becomes more difficult. Murphy (2015) implements this approximation by making some areas more salient than others in a process called 'Semantic Focusing.' Here, a selective blurring is applied to some regions in the image so that the user's attention is guided towards the sharp objects rather than the fuzzy ones.

2.4.2.2. Covering layers

Creating a clear distinction between foreground and background stands for an alternative to highlight and debase objects. Murphy (2015) gives to this technique the name of ‘Selective Brightening’, where a hierarchy between elements is defined by inducing haze. This method brings the appearance of a layer covering regions that do not provide user-relevant information, keeping areas aim to be highlighted in their original appearance.

3. Methodology

The methodology of this thesis is based on two main processes. The first stage concerns the calibration of an existing object recognition model capable of generating as well as labeling the object instances in a video recording. Subsequently, the segmented regions of the identified objects are modified through the implementation of image enhancement techniques, so that they can be visually highlighted or de-emphasized according to the information they provide. Finally, both techniques are used to rig up a model integrating the previous systems.

The following figure outlines the process described above.

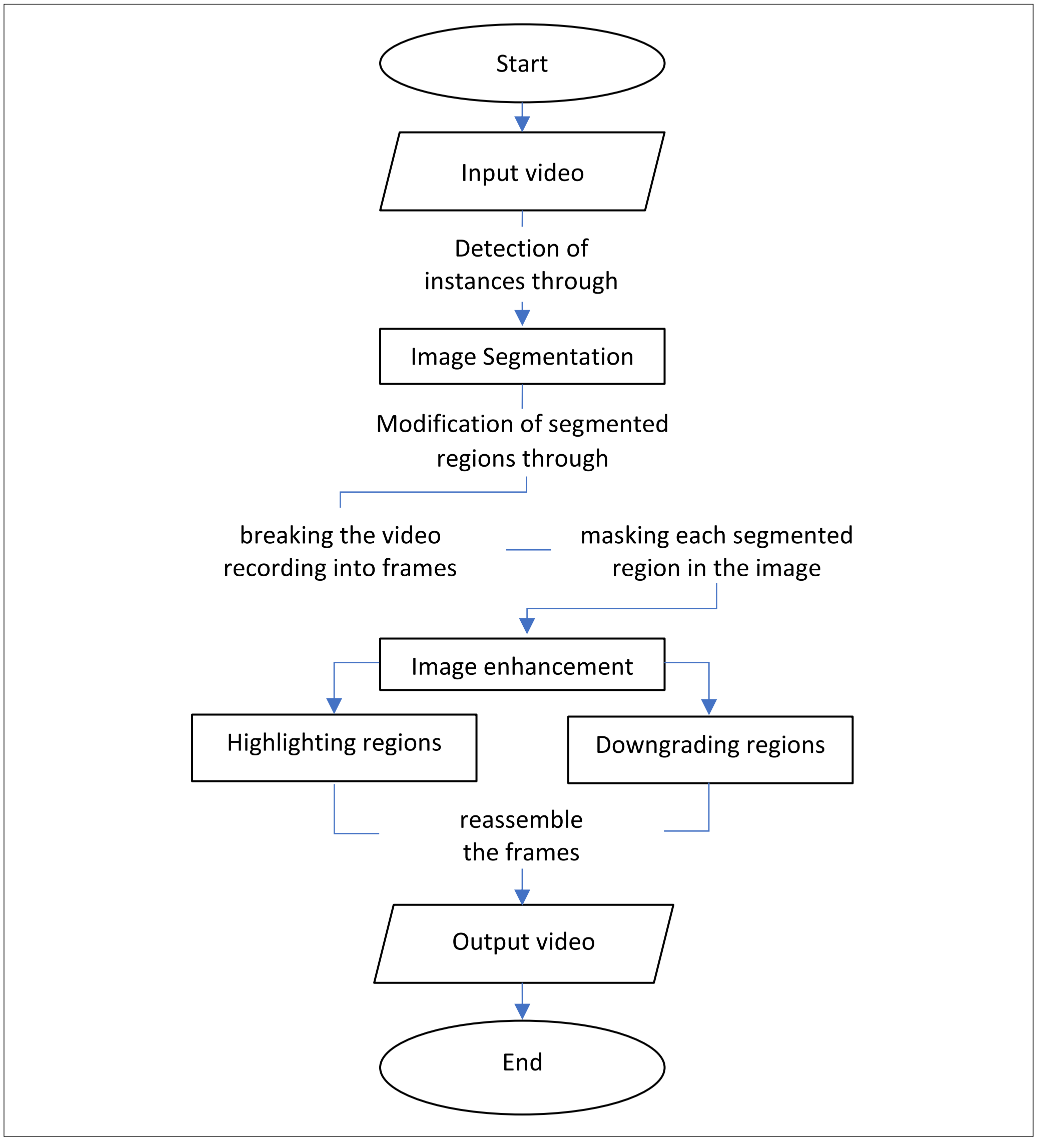

Start
Input video
Detection of instances through
Image Segmentation
Modification of segmented regions through
breaking the video recording into frames
masking each segmented region in the image
Image enhancement
Highlighting regions
Downgrading regions
reassemble the frames
Output video
End

Figure 8. Methodology workflow chart.

3.1. Object recognition model calibration

In order to identify the class of elements that are displayed during the video recording, an existing object recognition model needs to be calibrated. To accomplish this, some criteria is set for the selection of the model. Initially, the most important criterion refers to obtaining the instances of the objects. In this sense, given that the following stages of this methodology require manipulating the graphic properties of the objects individually, the model must be able to extract the object´s image borders.

Furthermore, since the input file corresponds to a video recording, the model needs to be capable of identifying and tagging objects in such a format. Similarly, the output file from the recognition process is required to have an outstanding image resolution, at least 720p (1,280x720), so that the boundaries of the segmented regions are clearly distinguishable to facilitate further processing.

Consequently, the analysis of the advantages and disadvantages of the object recognition approaches that are mentioned below, allowed to select the model that fit the methodological needs of this thesis. Likewise, this section shows the results of the implementation of the selected model and mentions some of its limitations.

3.1.1. Object recognition approaches

For this stage, the two fundamental approximations of object recognition were considered. Firstly, the object detection method was addressed with the use of the *YOLO* (You Only Look Once) algorithm, allowing recovering the spatial extension of the objects and enclosing it into a box along with its describing label class (Miksys et al., 2019). For the image segmentation approach, the *YOLACT* (You Only Look at Coefficient) algorithm was applied, generating geometries that cover only the regions that the objects occupy in the image (Hayder et al., 2017).

3.1.1.1. YOLO (You Only Look Once)

With regard to object detection, the model used was *YOLO* (You Only Look Once), whose latest version is *YOLOv4* (Redmon et al., 2016; Redmon and Farhadi, 2017; Nakahara et al., 2018; Redmon and Farhadi, 2018; Bochkovskiy et al., 2020). *YOLO* stands for a classifier performing object detection in both image and video files, whose main goal is to predict label classes and bounding boxes around objects. As can be seen in Figure 9, the spatial extent of each detected object is demarcated by a box, which is accompanied by the name of the category to which the object belongs.

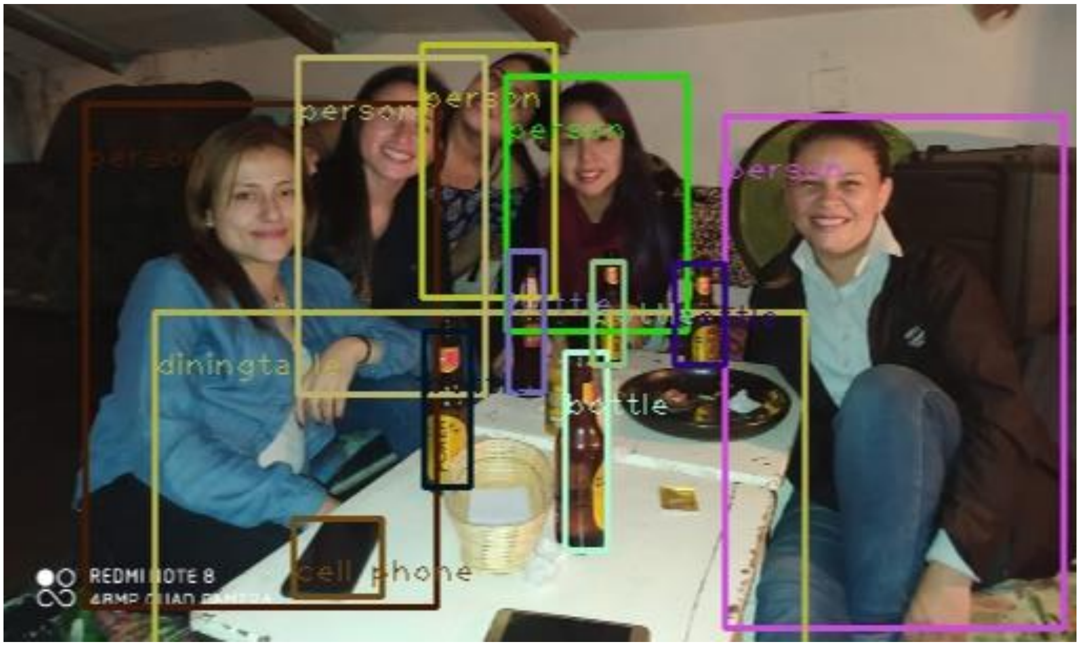


Figure 9. Object detection implementing YOLO (You Only Look Once).
Source: Author.

However, some drawbacks towards object detection are pointed out. Gould et al. (2009) acknowledge that bounding box-based object detector creates ambiguity when including pixels from different classes into the same boundaries, which reduces reliability over the elements identified. Likewise, Miksys et al. (2019) express that this method overlooks the physical features together with the posture of the objects, demanding additional steps to get their original shape.

3.1.1.2. YOLACT (You Only Look At Coeficient)

On the other hand, presented in 2019, *YOLACT* (You Only Look At Coefficient) is one of the most updated semantic segmentation models (Bolya et al., 2019). *YOLACT* represents a real-time semantic segmentation algorithm able to generate the instance of objects as well as their predicting bounding-box. Therefore, the method approached by YOLACT allows generating instance segmentation and object detection at the same time. As shown in Figure 10, the objects recognized by YOLACT are covered not only by the geometry of their shape (instance), but also by a container box that delimits their spatial extension.

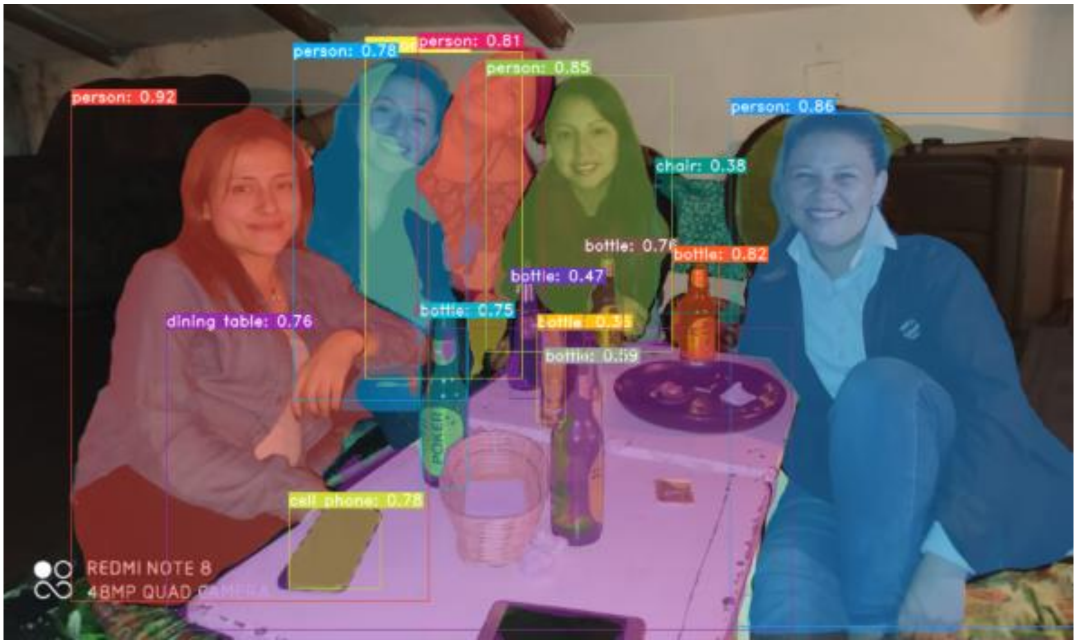


Figure 10. Instance segmentation employing YOLACT (You Only Look At Coefficient).
Source: Author.

Some constraints regarding instance segmentation models are related to the lack of spatial dimensions and differentiation between the delimitations of instances (Miksys et al., 2019). Notwithstanding, current algorithms, such as *YOLACT*, have overcome these limitations by adding bounding boxes to the recognition process, as illustrated in Figure 10. Moreover, Gould et al. (2009) estimate that a classification assigning each pixel to a unique class is an adequate approximation to produce useful outcomes and avoid bias when detecting objects.

3.1.1.3. Object recognition model selection

In this way, *YOLO* and *YOLACT* were the candidate algorithms applied to accomplish object recognition. *YOLO*, compared to *YOLACT*, requires less computational effort and a conventional GPU, bringing the possibility to be operated with no sophisticated hardware (Bochkovskiy et al., 2020). *YOLACT*, on the other hand, demands more hardware capabilities and advanced GPU. For this test data, *YOLO* was run on the *Anaconda* desktop environment and *YOLACT*, due to its complexity, was implemented on *GoogleColab*, a virtual machine powered by *GOOGLE* providing GPU for free.

Both approaches are capable of identifying and labeling objects; however, they produce different results. Object detection is useful to convey general information about the scene by demarcating square areas where the instance is contained, but there is no specification regarding the exact area covered by the objects in the image (see Figure 9). Conversely, instance segmentation provides precise object masks, preventing items from overlapping with items from different categories (see Figure 10) (Miksys et al., 2019).

Therefore, according to the objectives of this thesis, it is necessary to modify the regions occupied by the objects in the image, which requires delimiting their borders. In this sense, since object detection does not return the boundaries of the entity itself and overlooks specific features, instance segmentation, through *YOLACT*, is the approach addressed by the object recognition model in the present methodology.

3.1.2. Training dataset

YOLACT is designed in such a way that it is able to recognize the characteristics of objects by learning from images (Bolya et al., 2019). In this case, *COCO* (Common Objects in Context) is the name of the dataset feeding the algorithm, containing 2.5 million instances of 91 object classes in about 328,000 images depicting everyday situations (Lin et al., 2014). In each image filling the *COCO* dataset, an analog instance segmentation procedure took place, as shown in Figure 11, so that the object and its characteristics can be learned.

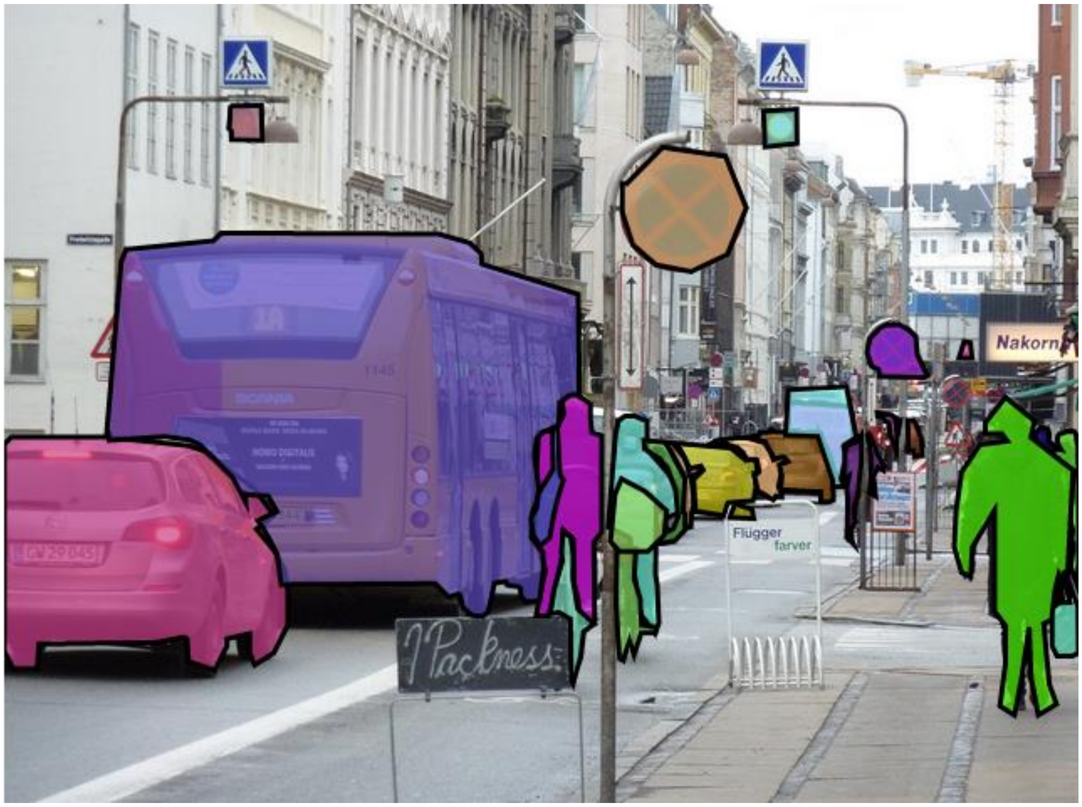


Figure 11. Analog segmentation of instances in COCO dataset.
Source: cocodataset.org.

Similarly, the following table shows the categories of objects that make up the COCO dataset, which also means that these are the objects recognizable by YOLACT.

Object	Super category	Object	Super category
person	person	cup	kitchen
bicycle	vehicle	fork	kitchen
car	vehicle	knife	kitchen
motorcycle	vehicle	spoon	kitchen
airplane	vehicle	bowl	kitchen
bus	vehicle	banana	food
train	vehicle	apple	food
truck	vehicle	sandwich	food
boat	vehicle	orange	food
traffic light	outdoor	broccoli	food
fire hydrant	outdoor	carrot	food
street sign	outdoor	hot dog	food
stop sign	outdoor	pizza	food
parking meter	outdoor	donut	food
bench	outdoor	cake	food
bird	animal	chair	furniture
cat	animal	couch	furniture
dog	animal	potted plant	furniture
horse	animal	bed	furniture
sheep	animal	mirror	furniture
cow	animal	dining table	furniture
elephant	animal	window	furniture
bear	animal	desk	furniture
zebra	animal	toilet	furniture
giraffe	animal	door	furniture
hat	accessory	tv	electronic
backpack	accessory	laptop	electronic
umbrella	accessory	mouse	electronic
shoe	accessory	remote	electronic
eye glasses	accessory	keyboard	electronic
handbag	accessory	cell phone	appliance
tie	accessory	microwave	appliance
suitcase	accessory	oven	appliance
frisbee	sports	toaster	appliance
skis	sports	sink	appliance
snowboard	sports	refrigerator	appliance
sports ball	sports	blender	appliance
kite	sports	book	indoor
baseball bat	sports	clock	indoor
baseball glove	sports	vase	indoor

skateboard	sports	scissors	indoor
surfboard	sports	teddy bear	indoor
tennis racket	sports	hair drier	indoor
bottle	kitchen	toothbrush	indoor
plate	kitchen	hairbrush	indoor
wine glass	kitchen		

Table 1. Categories of objects that make up the COCO dataset.
Source: Lin et al. (2014).

3.1.3. Calibration and application

The calibration and testing of *YOLACT* were executed in *GoogleColab*, which required a direct connection between the server and *Google Drive* to locate the algorithm code, input video, and output path. The input video used for this methodology was recorded in an exterior environment depicting urban surroundings. This shooting was recorded for fifteen (15) minutes following the direction of the sidewalk.

The categories identified by *YOLACT* in this sequence were:

Objects identified by *YOLACT*

Car	Truck	Traffic light
Traffic sign	Motorbike	Potted plant
Bicycle	Person	Dining table
Backpack	Stop sign	Chair
Clock	Bottle	Bench
Parking meter	Umbrella	Handbag
Bus	Fire hydrant	

Table 2. Categories of objects that were identified by YOLACT.

Figures 12, 13, 14, 15, and 16 display the result of the identification and segmentation processes in different sections of the video yielded by *YOLACT*. As noted, the instances of the objects are represented in different colors according to the category to which they belong. In this way, the segmented area of cars is green, people are represented in red, bikes in blue, traffic signs in grey, etc. The labels showing the categories of the objects are accompanied by a number indicating the probability that the object belongs to that class. Similarly, as part of *YOLACT* properties, not only instances of objects are segmented, the spatial extent of objects is also determined through the creation of bounding boxes.

On the other hand, it is important to note that, as shown in Table 1, the categories belonging to 'Buildings', 'Road', ‘Sky’, and 'Sidewalk' are not included in the training dataset of the algorithm. Therefore, as can be observed in the illustrations, these classes are neither recognized nor labelable by the object recognition model during this process.

The images also show that the resolution of the resulting video is high definition, i.e. 720p (1,280x720), which makes it clear to distinguish objects. Additionally, the result has a frame rate of 30 frames/second, creating a constant motion sensation. In this way, *YOLACT* responds to the requirements that were initially raised for the selection of the object recognition model.

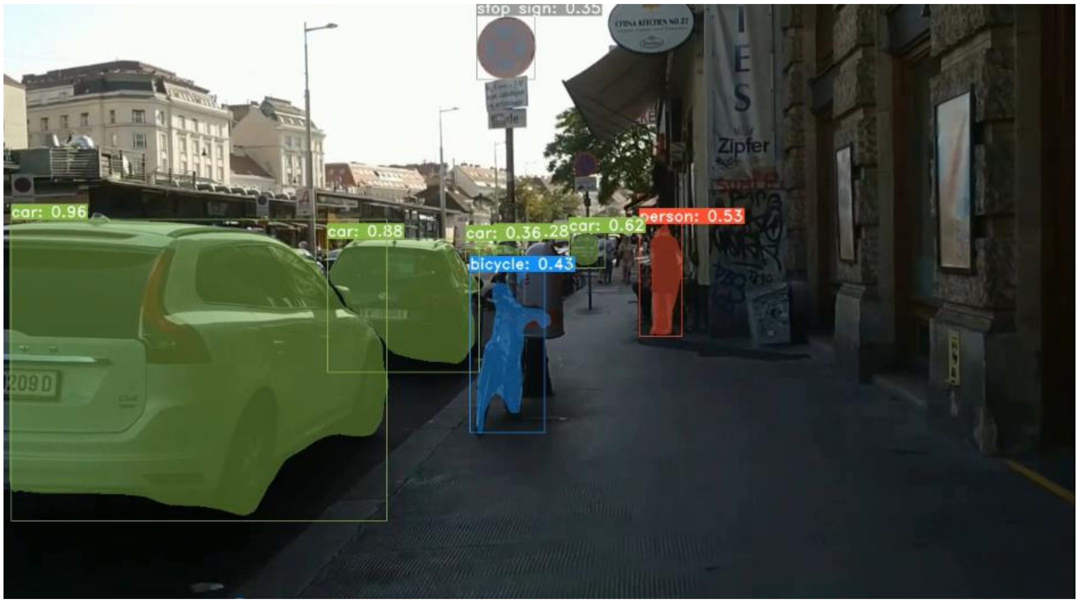


Figure 12. Scene 1 employing YOLACT for object recognition.

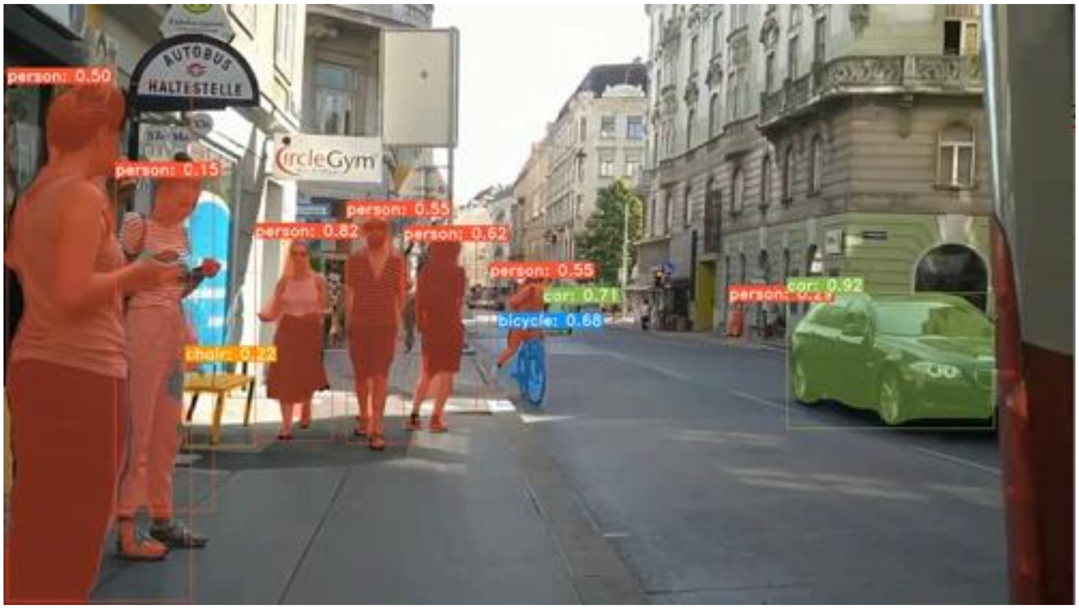


Figure 13. Scene 2 employing YOLACT for object recognition.

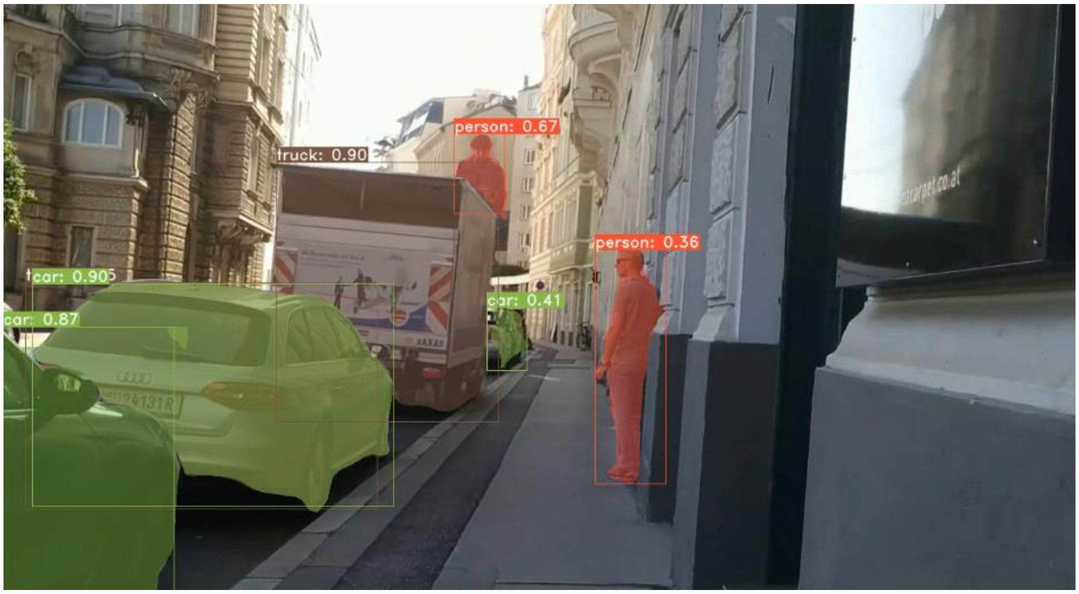


Figure 14. Scene 3 employing YOLACT for object recognition.

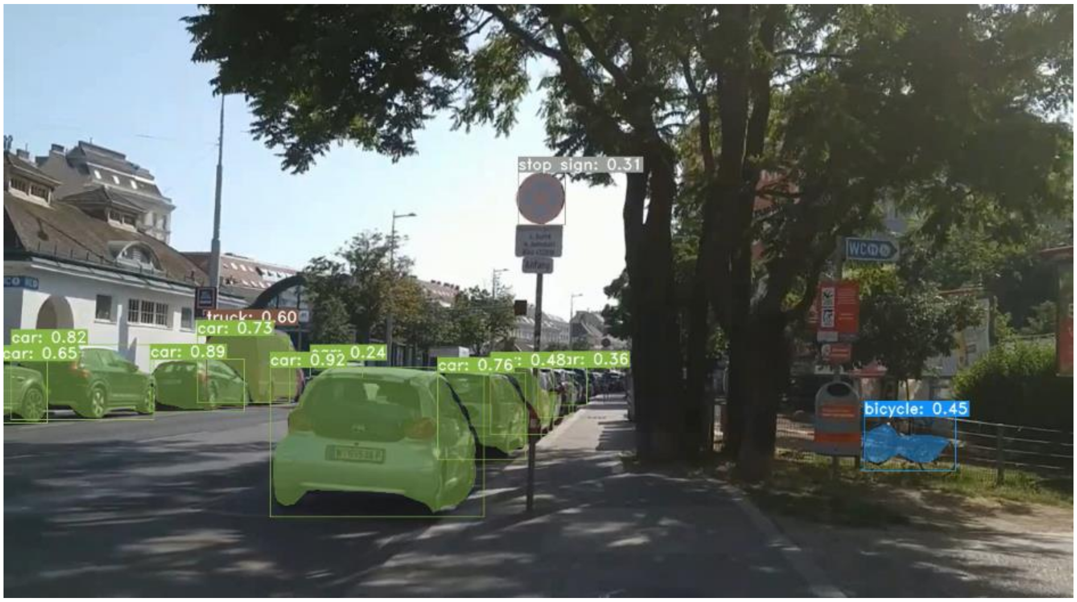


Figure 15. Scene 4 employing YOLACT for object recognition.

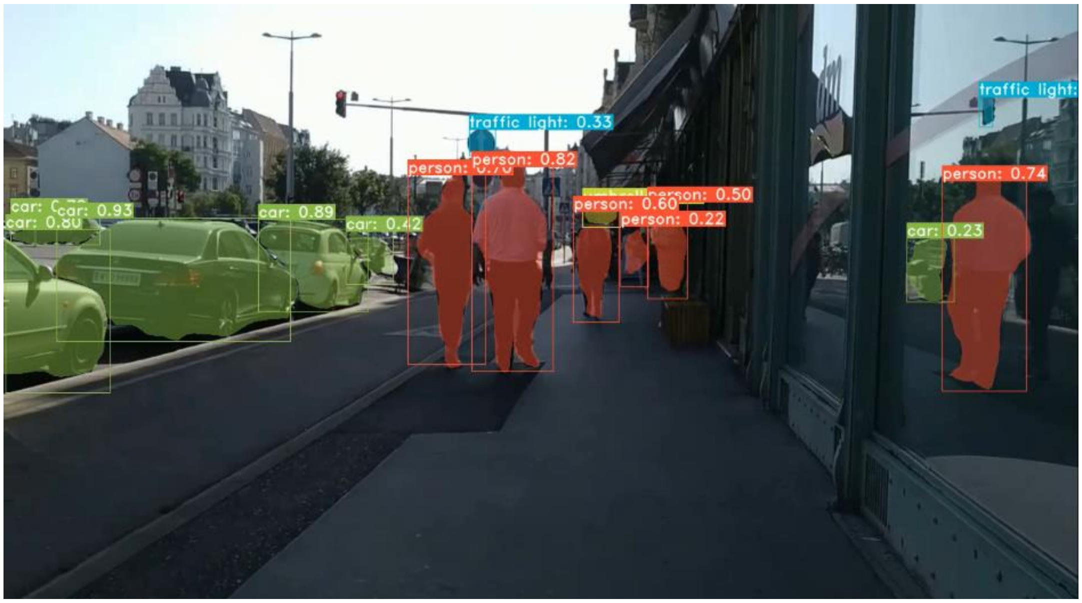
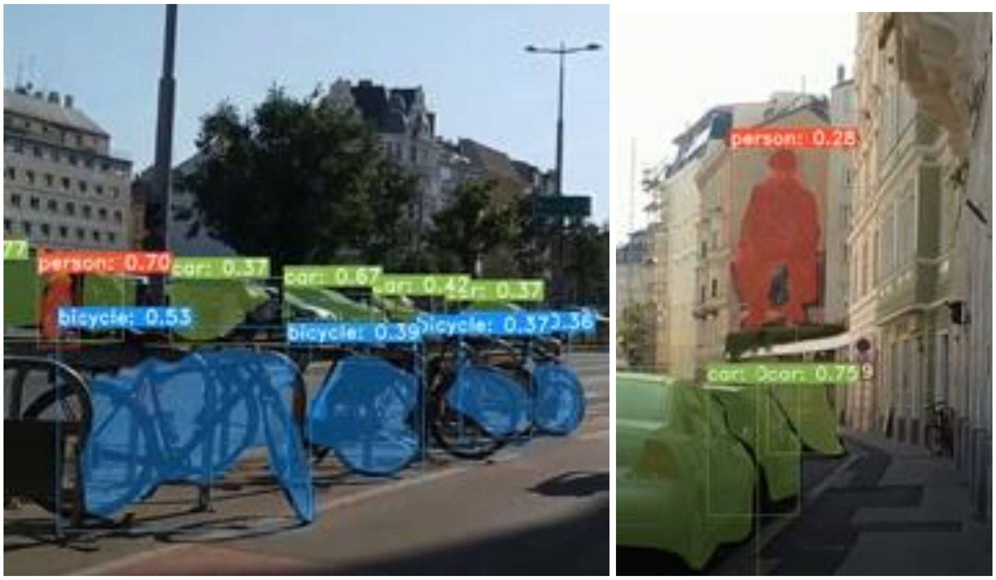


Figure 16. Scene 5 employing YOLACT for object recognition.

On the other hand, regarding the limitations of this model, it can be observed that, although the segmented areas are in most cases exact, some instances are not faithful to the objects they represent. This is especially the case with the bicycles, which, as can be seen in Figure 17(a), are not correctly segmented according to their shape. This limitation would prevent the object from being fully covered by the segmented region, making the image enhancement process inefficient, since not all details of the object are modified. In the same way, some objects were misclassified and consequently assigned to a label that did not correspond to them. For example, as illustrated in Figure 17(b), the painting on the wall was recognized by YOLACT as a person.



(a)Deformed instances (b) Misclassification

Figure 17. Limitations of YOLACT such as (a) deformed instances and (b) misclassification.

3.2. Image enhancement application

As pointed out in section 2.4., visual enhancement aims to improve the quality of representation of objects in an image so that the observer can recognize them more easily. At the same time, this process seeks to reorient the user's attention through the construction of a hierarchy where the regions providing non-useful information to accomplish a certain task are downgraded, while those regions standing for relevant details are emphasized.

Since this thesis aims to facilitate the navigation experience, the reasons for visually modifying the characteristics of the objects must be established. According to the classes that were identified by YOLACT (see Table 2), categories that offer guidance information and transit rules, such as traffic lights and traffic signs, are considered relevant to complete the navigation task (Esteban G., 2012). The opposite occurs with the remaining categories, such as people, cars, and bicycles, which do not provide the user with key elements to perform the mentioned undertaking.

In order to determine how distracting these elements might be in an urban environment, a user test was conducted using *Google Forms*. This user test was thought to serve as an indicator of how attention can be directed to certain objects. Additionally, a digital element was added to the video recording used in this test, so that a mixed reality context could be recreated, as shown in Figure 18.

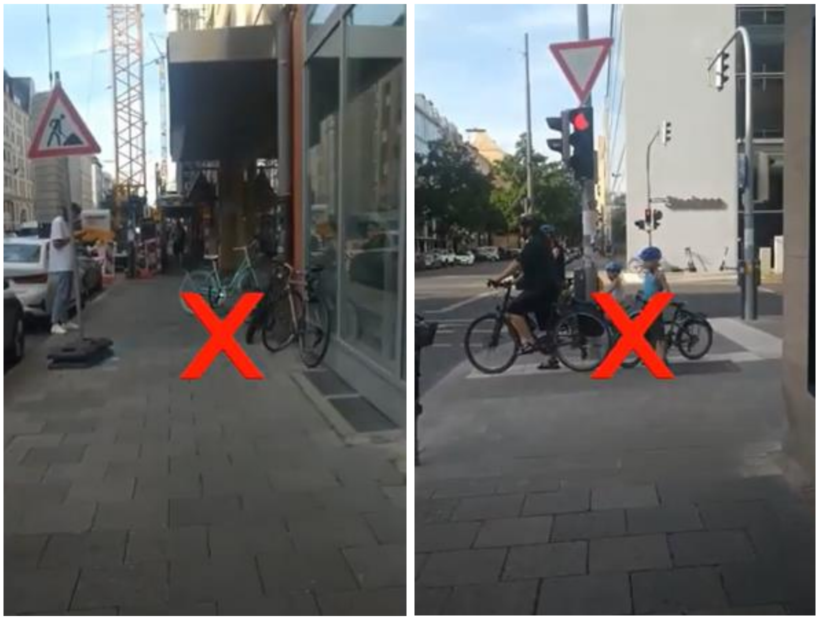
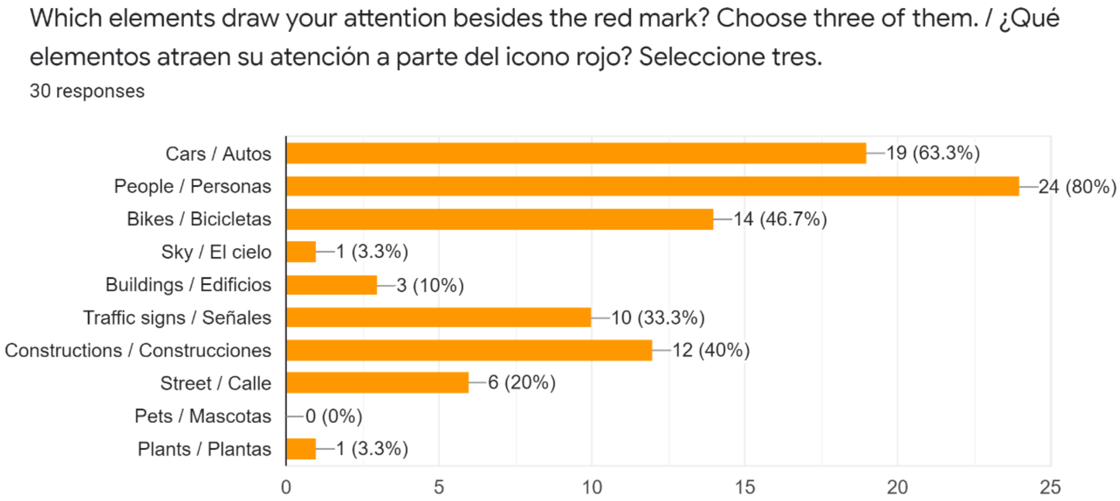


Figure 18. Scene to determine striking elements.

A total of 30 participants, university students, took part in this test. They were asked to focus their attention on the digital element and then select the three elements that were most striking to them (the video recording used in this test is available at https://www.youtube.com/watch?v=Saqb3c5EQds, last retrieved October 1, 2020).

Taken directly from *Google Forms*, Figure 19 illustrates the responses for the question stated. Here, the elements that were considered to be the most conspicuous were the ones moving along the video: people, cars, and bikes. Although the construction event was portrayed for a very short segment during the shooting, participants also rated it as striking. Traffic signals, which were deployed in several scenes, were observed to a lesser extent. On the other hand, pets, plants, buildings and the sky, were rated as the elements which got less attention.



Which elements draw your attention besides the red mark? Choose three of them. / ¿Qué elementos atraen su atención a parte del icono rojo? Seleccione tres.

30 responses

Figure 19. Responses regarding striking elements during the video.
Source: Google Forms.

Therefore, this stage addresses two methods to achieve visual enhancement: highlighting and downgrading. It is important to mention that the information provided by traffic signs has a greater relevance to complete the navigation task compared to that provided by objects such as cars, people, and bikes. Similarly, as can be seen in Figure 19, traffic signals were less noticeable to users than other categories of objects. Thus, highlighting techniques will be applied to categories of objects that provide the user with information essential to navigate correctly, while objects that are more noticeable but transmit less meaningful information will be downgraded.

3.2.1. BLENDER 2.83.0

The visual improvement techniques were implemented in a section of the video recording where the instances of the objects were detected and extracted. The section corresponds to a one-minute piece of footage composed of 1,818 frames. In each frame, the geometries of objects were accentuated or degraded according to the information they provided.

The tool selected to carry out the highlighting and downgrading operations was *BLENDER* version *2.83.0*, which stands for an open-source software used for generating digital graphics. *BLENDER* is able to break up the footage frame by frame and adjust the attributes of each scene. Due to the requirement to work on the very specific segmented areas, *BLENDER* brought the possibility to focalize on these regions through the process of masking, which consists of creating geometric shapes that are identical to the figures being superimposed.

The masking process was performed manually on each of the 1,818 frames that make up the video sequence, so that the segmented regions could be carefully delimited in order to ensure that changes to the pixel properties affected only the instances of the identified objects. As Figure 20 shows, the masks were created from a series of points surrounding the segmented region, allowing to track the displacement of each object frame by frame. It is also worth mentioning that the masks were elaborated from the frame in which *YOLACT* recognized each particular object.

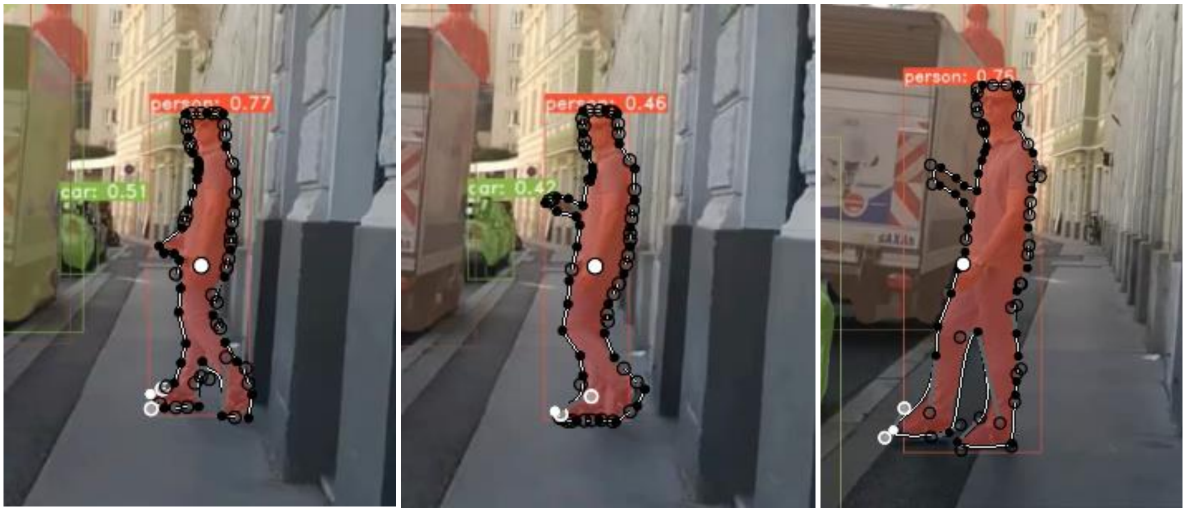
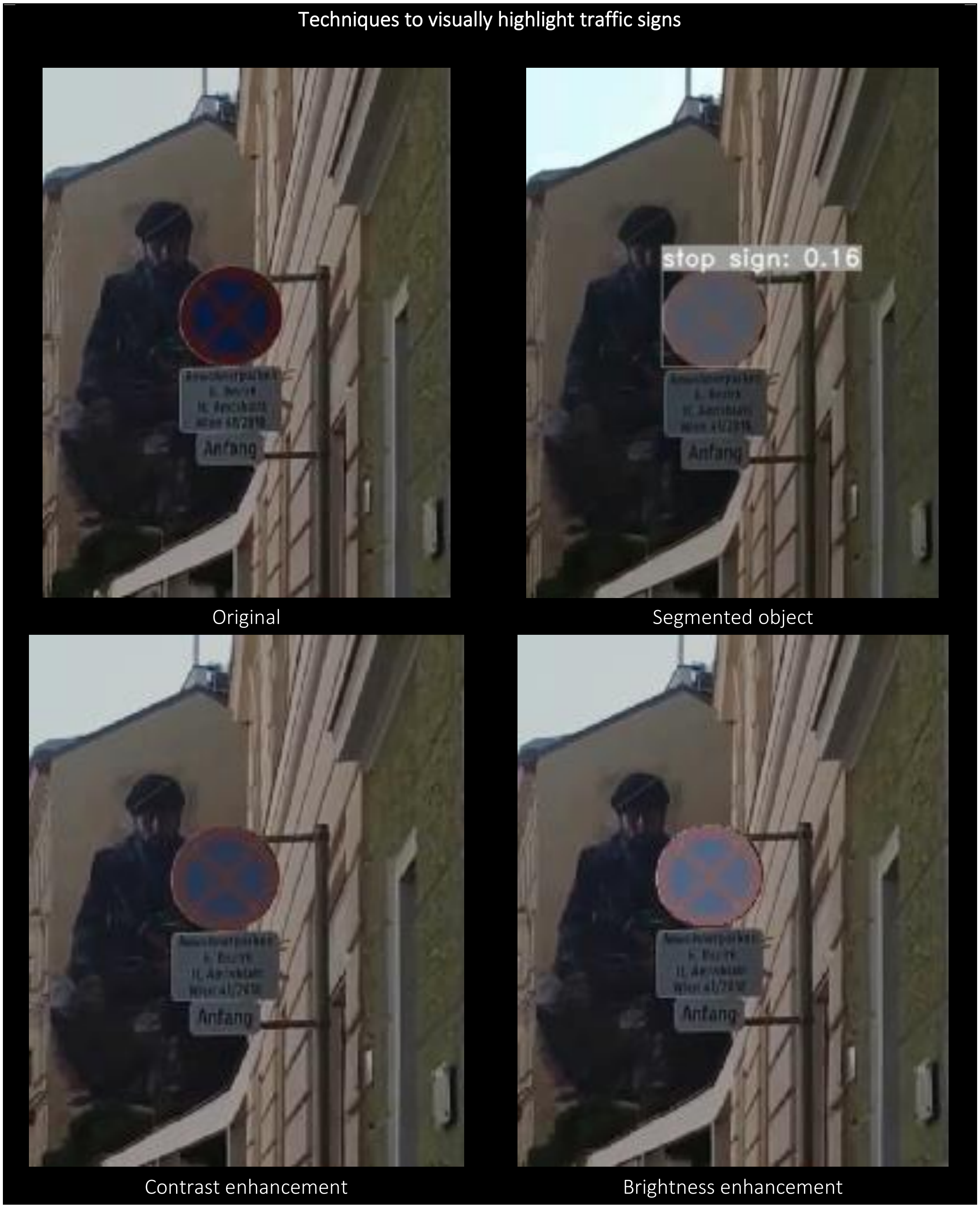


Figure 20. Elaboration of masks in BLENDER.

3.2.1.1. Object accentuation

The process of accentuating the attributes of the objects is known as highlighting (Maini and Aggarwal, 2010). As indicated in section 2.4.1., contrast, brightness, and saturation stand for the main techniques to modify pixel properties to emphasize their attributes (Kaur, 2013; Saleem et al., 2012; Yoon et al., 2009). In this thesis, these visual enhancement methods are applied to the traffic signs and traffic lights, which are thought of as pictorial tools that provide information to facilitate navigation and bring safety to commuters (Esteban G., 2012).

Table 3 deploys the visualization of the same traffic sign through the employment of the previous image enhancement options along with the segmentation output. It can be perceived that the modification of contrast makes a moderately better distinction of details; however, they are not clear enough. Brightness, on the other hand, gives a haze effect that spotlights the sign from the general scene but covers up its inner attributes. Conversely, saturation enhances the characteristics of the object itself and creates an effect of accentuation compared to other parts of the image.



Techniques to visually highlight traffic signs

Original

Segmented object

Contrast enhancement

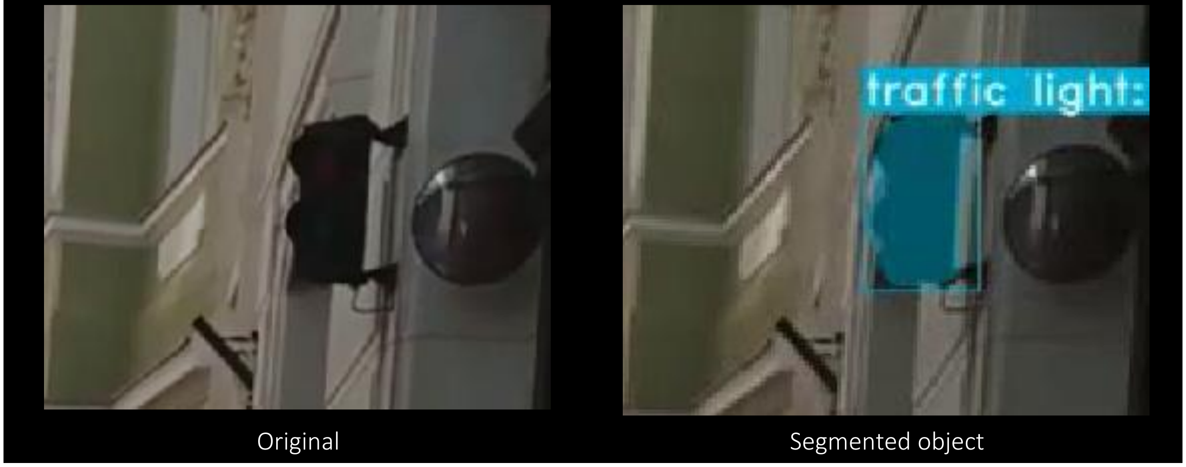
Brightness enhancement

Saturation enhancement

Table 3. Visual enhancement techniques applied to traffic signs.

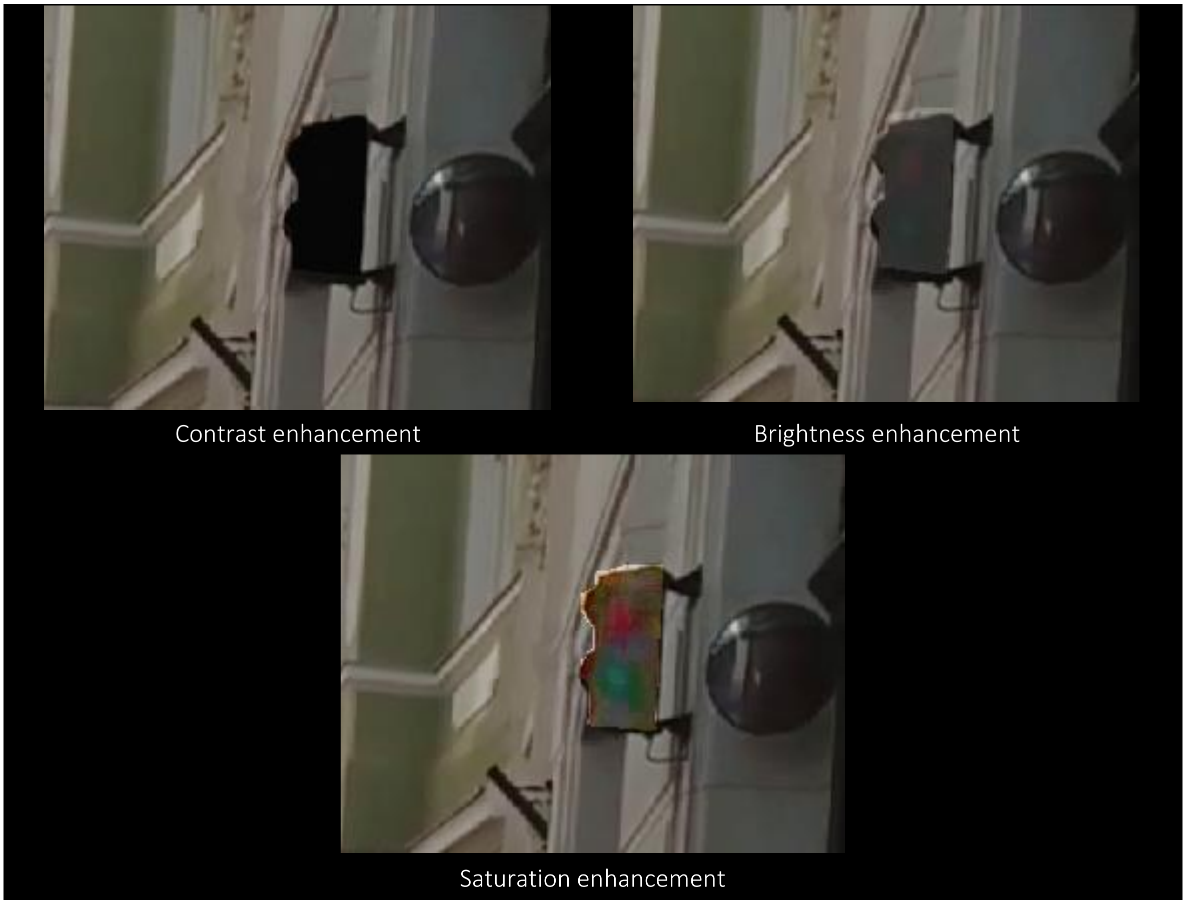
In addition to traffic signs, Table 4 shows the application of contrast, brightness, and saturation enhancement on traffic lights. In this case, the effect of contrast produces a darker appearance, and, similar to Table 3, brightness hides the object behind a semi-transparent white layer. With regard to saturation, although the initial tone of the traffic light is altered, this improvement allows identifying the green and red lights that are not distinguishable in the original picture.

Techniques to visually highlight traffic lights



Original

Segmented object



Contrast enhancement

Brightness enhancement

Saturation enhancement

Table 4. Visual enhancement techniques applied to traffic lights.

3.2.1.2. Object downgrading

The downgrading techniques aim at degrading the original properties of the object to make it look like part of the background as well as redirecting the user's attention to elements providing more relevant information (Murphy, 2015). Table 5 addresses the methods discussed in section 2.4.2., whereby the decreased recognition of regions in the image is achieved by either blurring the object or hiding its features.

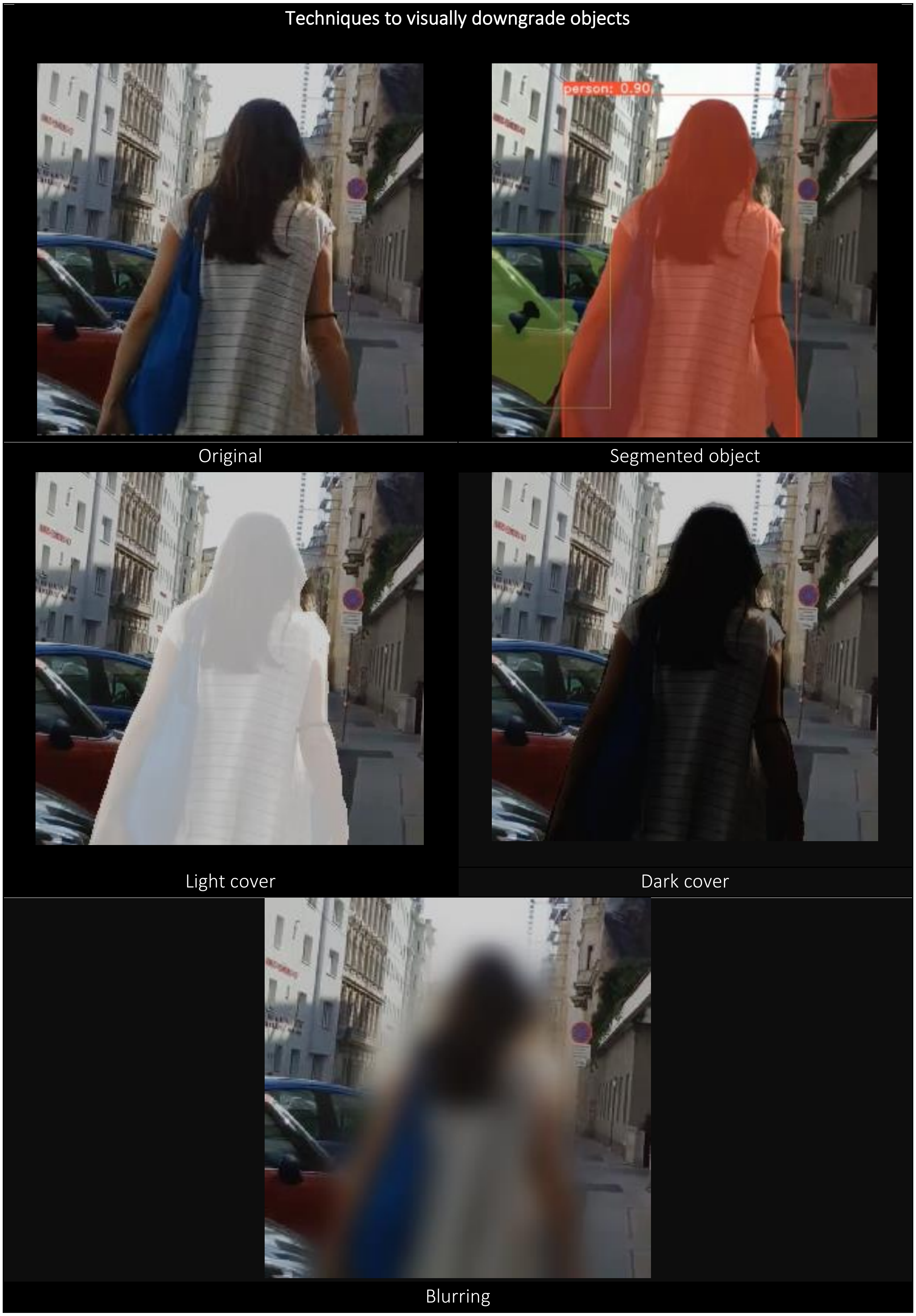


Table 5. Image debasing methods.

As can be seen in Table 5, the use of the light cover creates a haze effect on the object but keeps its characteristics clearly visible. The dark cover hides the object, but, due to its contrast, creates a predominant black region in the image. Blurring, on the other hand, distorts the object's features without completely occulting it or creating a salient region. Therefore, blurring is considered the most appropriate technique for downgrading object properties. In table 6 some categories of objects are blurred based on their segmented areas.

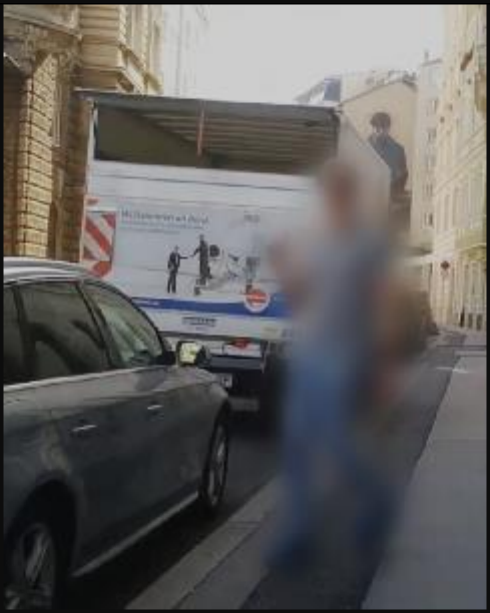
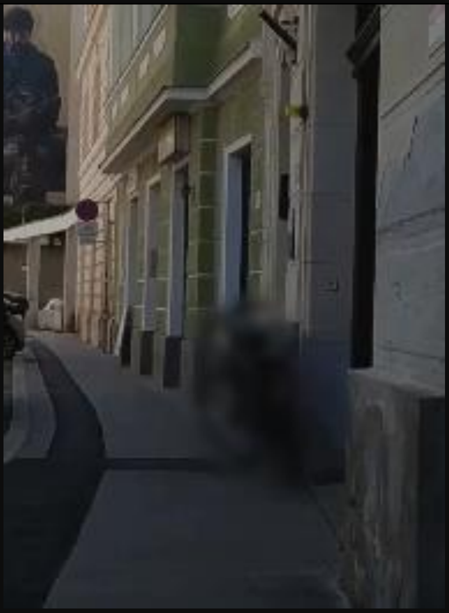
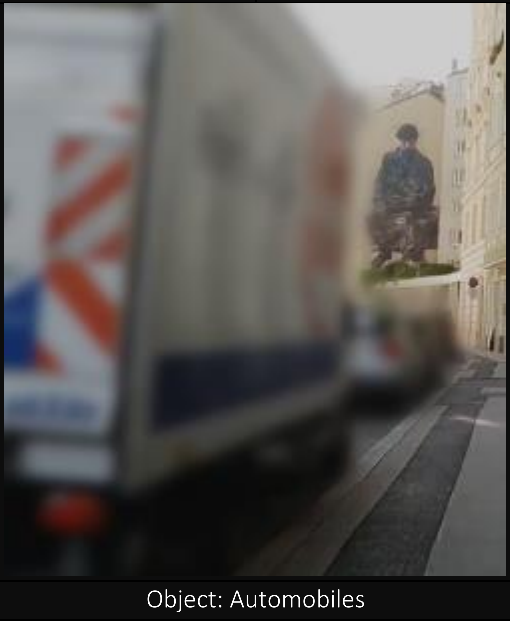
Blurred objects	
	
Object: Person	Object: Bicycle
	
Object: Automobiles	

Table 6. Blurring applied to different categories of objects.

3.3. Implementation

The integration of the image segmentation model together with the visual enhancement methods were applied to a segment of the original input video, which was composed of 1,818 frames. As addressed in section 3.1. and 3.2., the elements identified by the object recognition model were visually enhanced through techniques of blurring and modification of the saturation level, so that their visual properties could be highlighted and downgraded.

Once the frames were fully processed, each frame was reassembled, forming a resulting one-minute video recording (available at https://youtu.be/hETJwk4pUwU, last retrieved October 1, 2020). This outcome had a frame sequence similar to the original video recording, i.e. 30 frames/second, and also a resolution of 720p (1,280x720).

The following images show scenes resulting from the integration of the image segmentation algorithm and the visual enhancement methods. This video recording displays the path through the pedestrian area from a pedestrian's point of view. As can be seen in the images, only the regions occupied by the objects were visually highlighted or downgraded. Similarly, during video playback, the displacement of these regions tracked the movement of the objects.

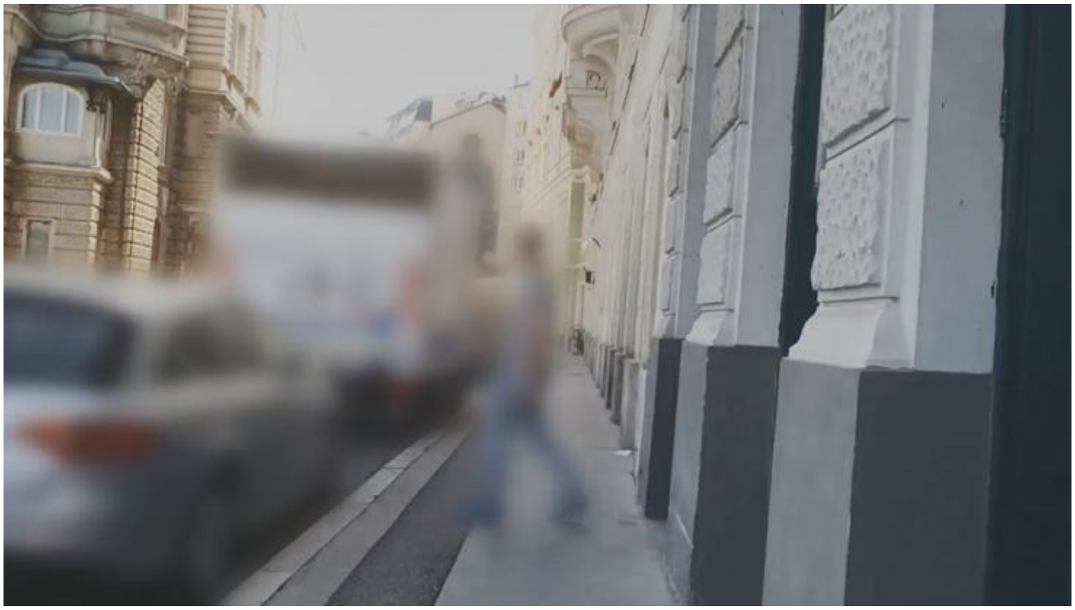


Figure 21. Scene 1 from the integration of models.

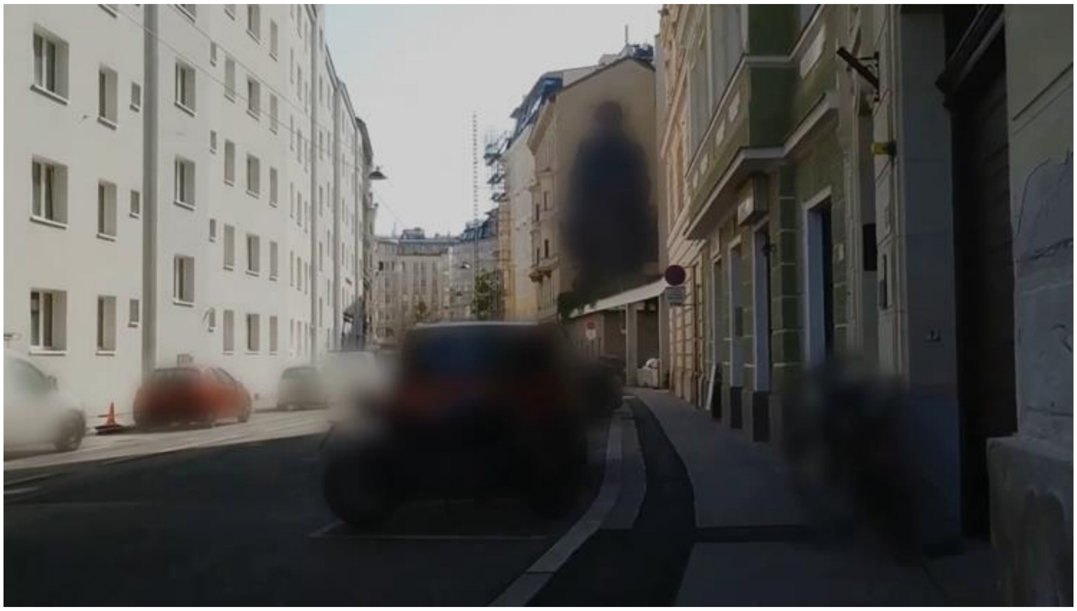


Figure 22. Scene 2 from the integration of models.

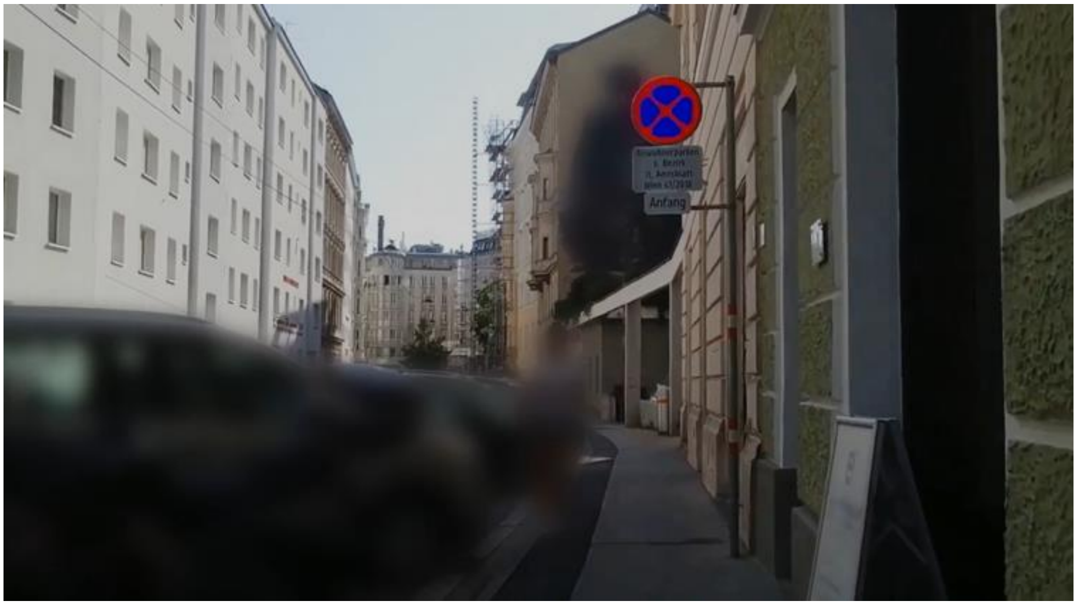


Figure 23. Scene 3 from the integration of models.

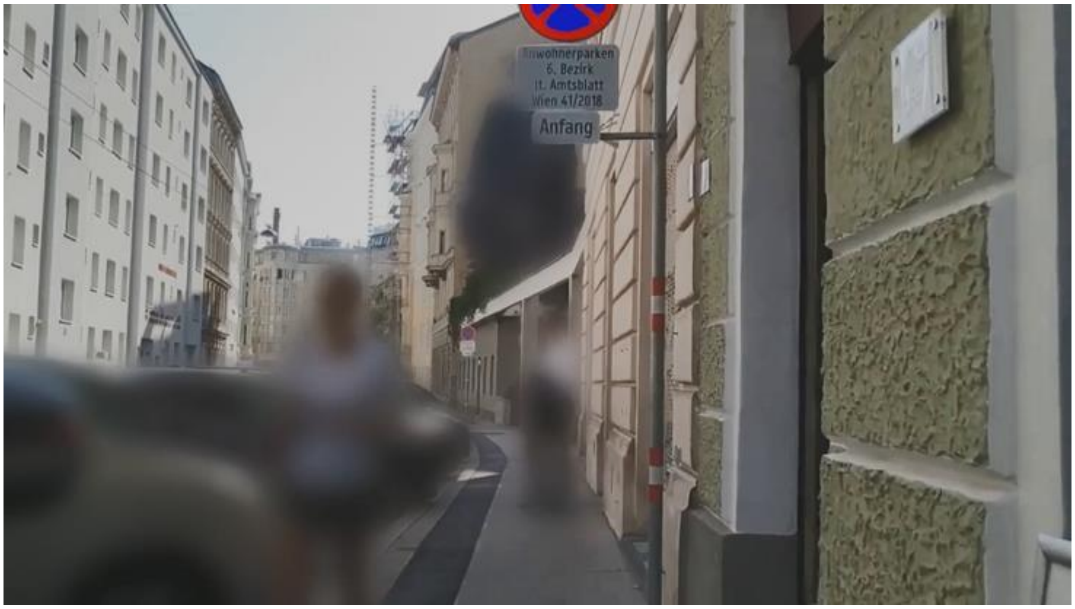


Figure 24. Scene 4 from the integration of models.

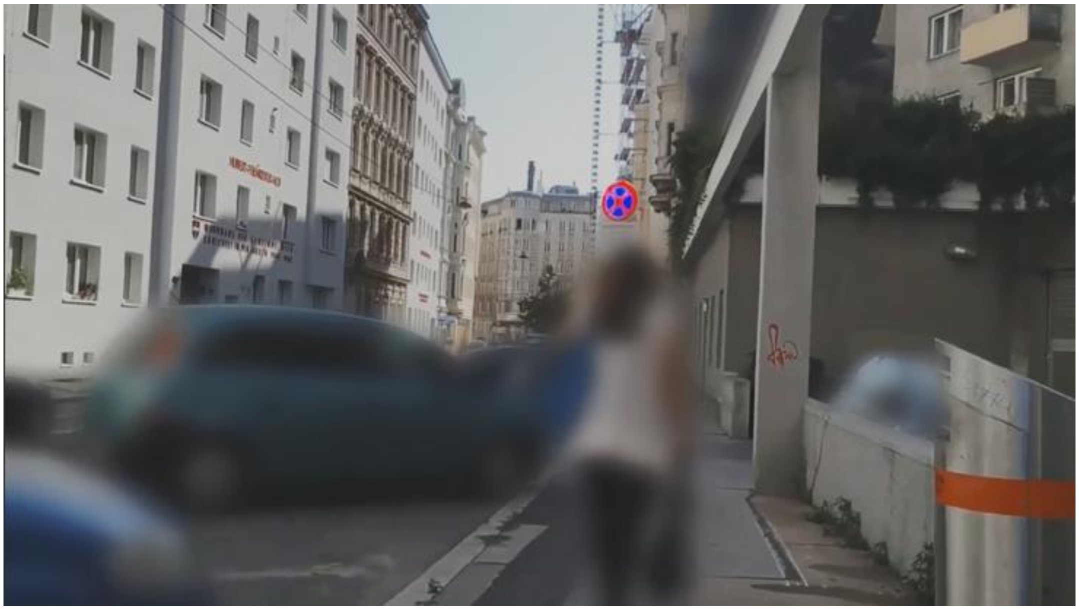


Figure 25. Scene 5 from the integration of models.

4. User tests

According to Tacca (2011), a series of stages lead from perception to cognition. In the case of visual perception, the early stages compromise the segregation of objects from their background, detection of borders and identification of features. Subsequent processes take place in memory, where the perceived objects are stored and compared to other representations that have been previously observed.

In this way, as Tacca (2011) states, the initial steps of visual perception occur through the senses and, once the subject is aware of the information received, the process takes place in the field of consciousness. For this reason, the improvement of graphic representations is intended to generate a stimulus, so that the recognition of certain objects is facilitated.

Therefore, in order to estimate the impact that the proposed methodology had on the change of perception of the users when they are simulating a navigation experience, two user tests were conducted. In the first test, the video that was displayed did not suffer any modification. The second test used the same recording, but certain objects were highlighted, and others downgraded depending on their usefulness to navigate. In both videos, an X-shaped rotating three-dimensional object was added using the Windows 10 video editing tool. The embedded effect was arranged to create a mixed reality environment, in which the digital object traces the trajectory of the route through the sidewalk.

4.1. User test I: Identification of striking objects

The first approach to the case study was to identify the classes of objects that were most striking to users through the playback of a video recording showing a traditional urban landscape for one minute. This task was accomplished by employing a questionnaire created in Google Forms (available at https://forms.gle/efs4AHkXKwbKRT8P8, last retrieved October 1, 2020), in which fifty (50) participants, including students and university professors from several fields, took part. Here, users were asked to look at a digital element superimposed on the video recording, as depicted in Figure 26, and then select some items according to the following two multiple-choice questions:

1) Besides the red icon, which elements draw your attention the most?
2) Which elements call your attention the least?

For both questions, participants were requested to select three of the elements shown in Table 7, so that a pattern of the elements that were most striking to them could be established.

List of items to select		
Cars	Sky	Constructions
Traffic signs	Buildings	Street
Bikes	People	Others

Table 7. List of items to select during the user tests.

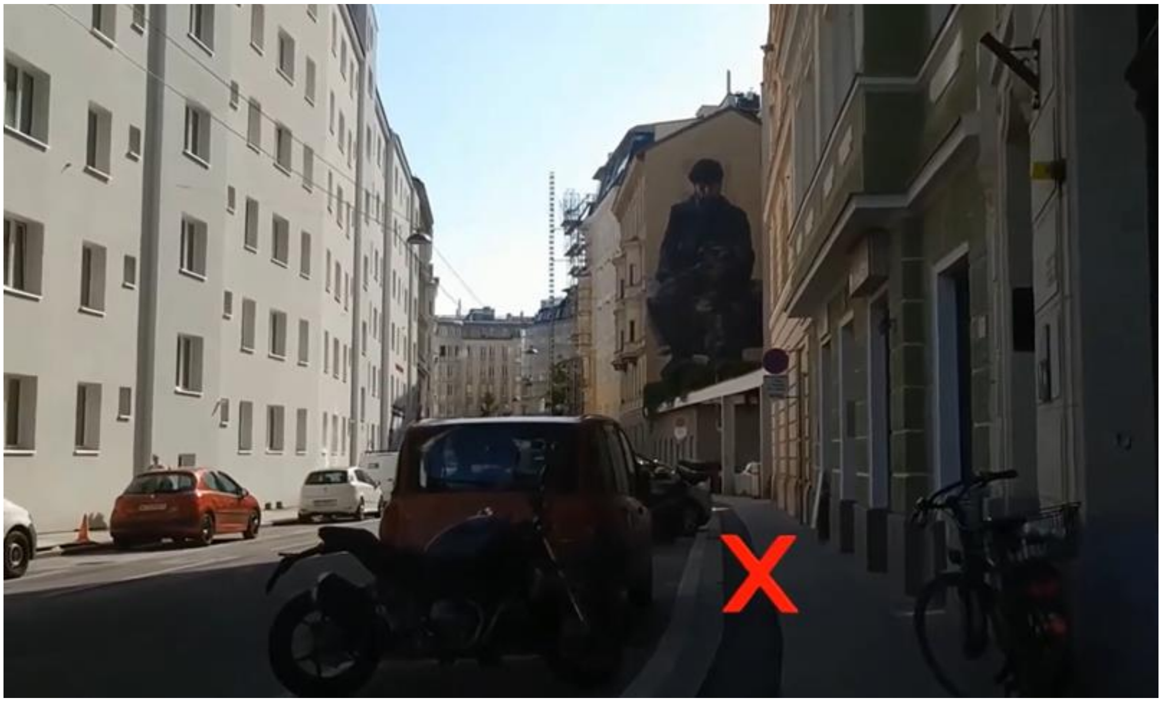


Figure 26. Scene from the first user test video.

According to the responses given on *Google Forms*, Figure 27 illustrates how eye-catching every class of items on the list was. The categories assessed to be most distinctive besides the digital item were People, Cars, and Buildings. On the other hand, Sky, Traffic signs, and Bikes were considered as inconspicuous elements. Although the classes of Construction and Street were portrayed during the whole shooting, participants rated them as less conspicuous. Lastly, in the category Other, users mentioned the mural in the background.

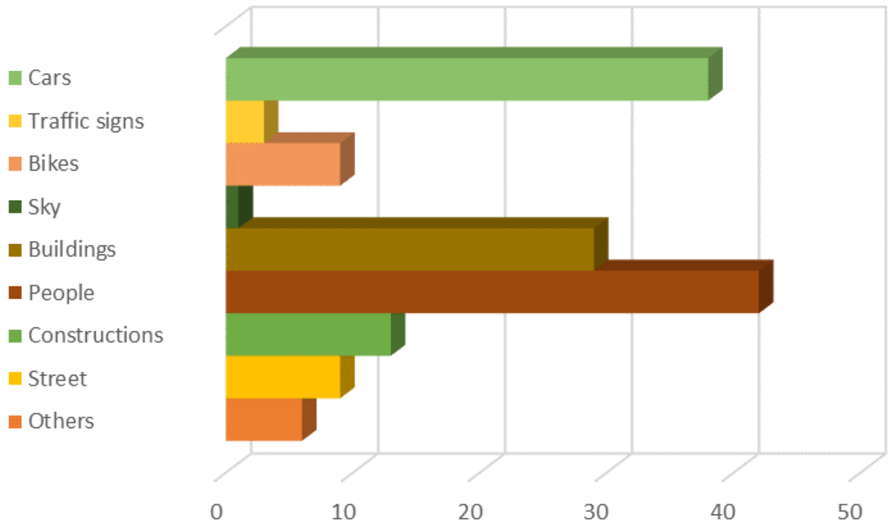


Figure 27. Responses regarding the most striking objects for the first user test.

Replies to the second question match with the answers shown in Figure 27. In this case, participants determined that Traffic signs and Sky were the classes that kept their attention from the red mark the least. Additionally, Bikes and Street ranked as low distractive items. Conversely, People, Cars, Buildings as well as Constructions were more noticeable, as exposed in Figure 28.

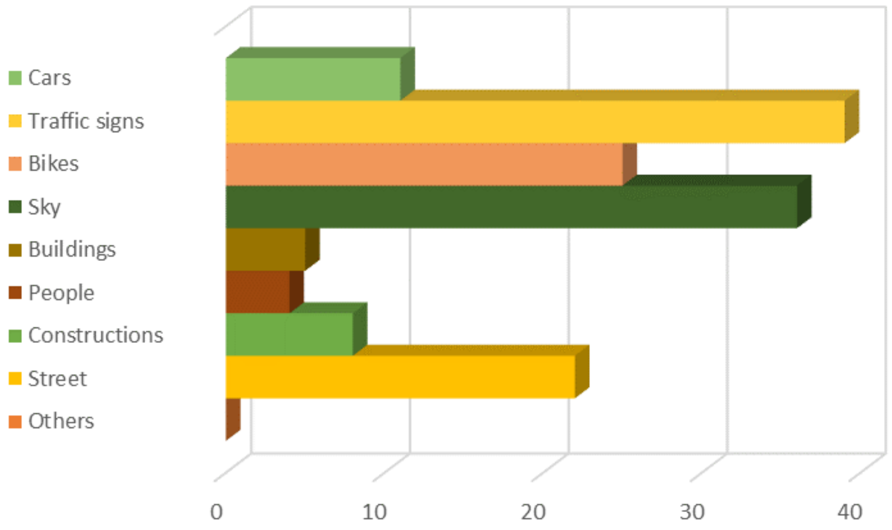


Figure 28. Responses regarding the least striking objects for the first user test.

4.2. User test II: Identification of striking objects applying image enhancement

To evidence how users' attention is affected by highlighting and degrading certain objects, a second user test was performed. The video recording, in this case, was the result of detecting some categories of objects and applying image enhancement methods on them (see Figure 29). The questionnaire and the list of options were similar to those deployed in section 4.1. Similarly, this survey was shared and filled out through *Google Forms* (available on https://forms.gle/XmUwWDt44VLG81HGA, last retrieved October 1, 2020), which was also answered by 50 participants. Nevertheless, users that participated in this test were not the same as those who took part in the first test to avoid any kind of bias.

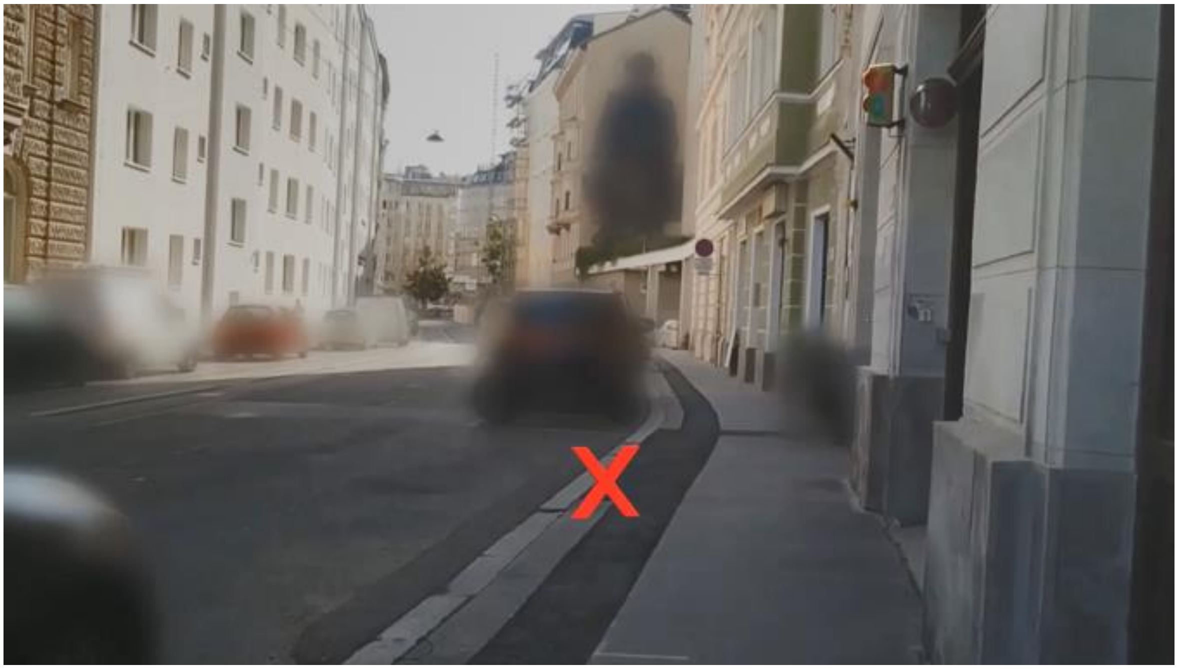


Figure 29. Scene from the second user test.

Figure 30 presents the responses of the first question with respect to the elements that most caught the users' attention besides the red X-shaped icon. Cars, People and Buildings were the most predominant categories of objects, followed by Traffic Signs. The Street class was evaluated by less than half of the participants as striking. As for categories Bikes and Sky, these attracted the attention of the participants to a lesser extent. In the Others category, one participant mentioned pavement.

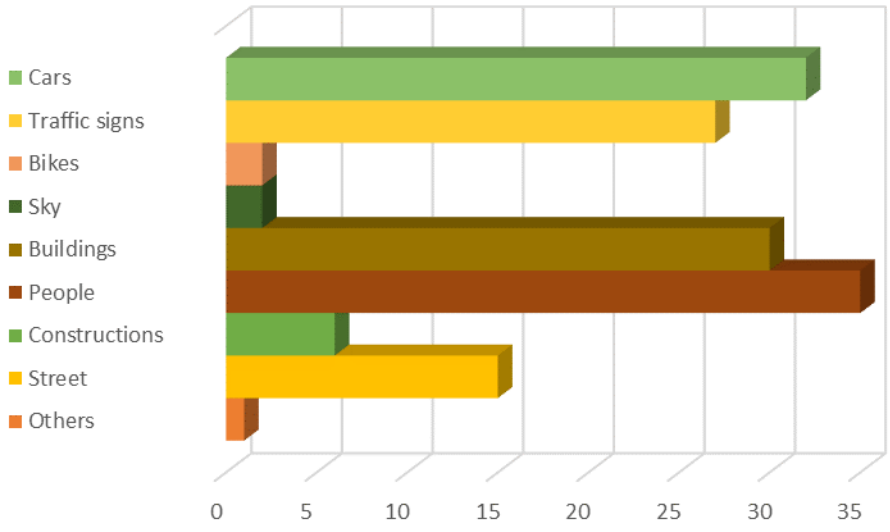


Figure 30. Responses regarding the most striking objects for the second user test.

With regard to the second question, Figure 31 illustrates that the sky was the class that attracted the least attention, followed by the categories of Bikes, Street and Constructions, which is equivalent to the result obtained in Figure 30. Cars, People, Buildings and Traffic Signs were considered by users to be more noticeable object categories.

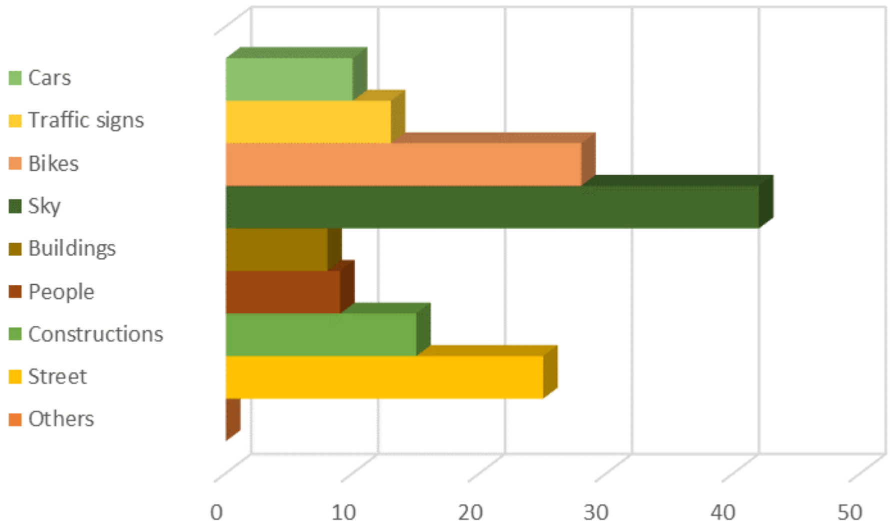


Figure 31. Responses regarding the least striking objects for the second user test.

5. Results

As mentioned in section 4, the identification of objects through the observation of their uniqueness and characteristics corresponds to the field of perception (Tacca, 2011). As a consequence, the two user tests that were conducted attempted to establish whether through the generation of a visual stimulus certain categories of objects were more noticeable to users and, at the same time, determine the effectiveness of the proposed methodology of this thesis.

Thereby, in this chapter the results collected from both user tests are explored, analyzed, and discussed. Similarly, the answers to the questions initially posed in the introductory part of the document are addressed. Finally, some recommendations are given for the difficulties encountered during the development of the objectives as well as the limitations that user tests had.

5.1. User tests findings

As stated in section 3.2, the category that was visually highlighted corresponded to Traffic signs, while the categories of Cars, Bikes and People were downgraded. In conformity with the results shown in the two user tests (see sections 4.1 and 4.2), there was a change in the users' perception of the object categories displayed. To make a clear distinction between the results gathered, Figure 32 compares the participants' evaluation of the elements that most attracted their attention with and without applying the methodology proposed in this thesis.

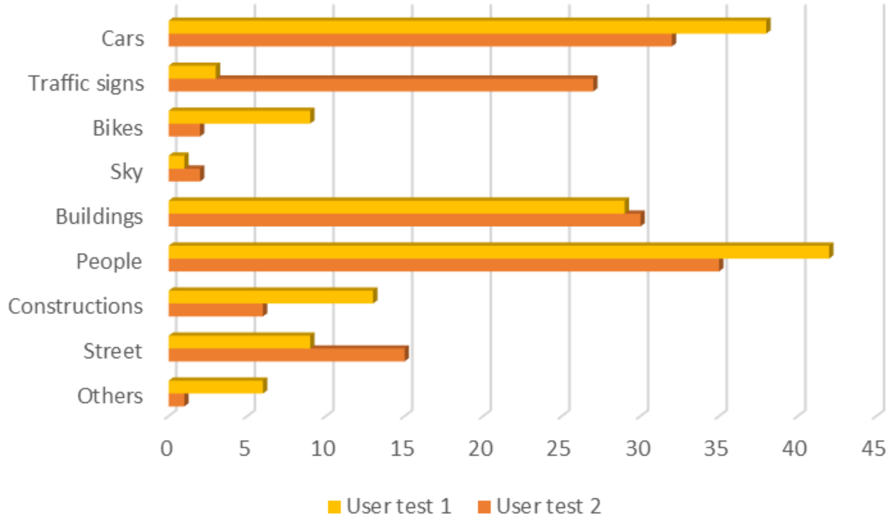


Figure 32. Comparison between User tests 1 and 2.

The horizontal axis of the graph that stands for the number of participants selecting the objects as distractive shows that the blurring effect applied to Cars, People, and Bikes categories did not have a significant impact in trying to keep attention from them. Although there was a slight decrease for all (nearly 15% for Cars and People, and roughly 80% for Bikes), Cars and People remained the most noticeable classes.

On the other hand, the highlighting of characteristics of Traffic signs had a greater incidence in user perception. According to Figure 32, Traffic signs gained nine times more notoriety, being one of the most prominent categories in the video recording used in the second user test.

Consequently, the evidence showed that de-emphasizing visual features of objects by blurring diminished the attention users pay to them, but not in a way that they lack relevance and cease to be striking. Conversely, the accentuation of objects through the modification of saturation values increased their visibility, allowing users' attention to be redirected to specific regions of the image.

5.2. Discussion

Through the application of the user tests, it was possible to determine that users' attention could be redirected when the visual features of the displayed objects are detected and modified. It is important to mention that the stimulus in users did not have the same response when it came to the use of highlighting and downgrading techniques. As shown in section 5.1., the visual intensification of objects had a much greater influence compared to the process of de-emphasizing, indicating less effectiveness of redirecting the users' attention by degrading the properties of objects.

So why did objects continue to be striking? In contrast to approaches considered to lessen attention, in which motionless images were used (see Murphy, 2015), the techniques implemented in this thesis were applied to video format files, setting different conditions for the objects displayed. In this manner, the continuous sequence of images introduced movement to objects, which represented a distracting factor. Similarly, the blurred regions did not form the entire background, making their displacement during the sequence noticeable to the observer.

However, as stated by Kaur (2013) and Maini and Aggarwal (2010), due to the complexity of the human vision mechanism, the changes in perception are subjective, finding different reactions to certain stimuli. Thus, although it was not the purpose of this research, categorical reasons cannot be established to explain users' responses to some approaches used in this thesis. Notwithstanding, this context opens up the possibility to explore in detail the relationship between certain visual effects and the impact they generate on human vision.

Consequently, the use of an image segmentation model along with visual enhancement techniques made it possible to identify as well as enhance and degrade the visual characteristics of objects in a video. Aligning with related research, some approaches have already addressed navigation in Mixed Reality (Knutsson and Georgsson, 2019; Carmigniani et al., 2011; Mehdi et al., 2020), and this work contributes by providing a greater context about the environment, facilitating the development of the task.

5.3. Development of research questions

Through the development of the objectives of this thesis, it was possible to give answers to the research questions stated in section 1.2.2. The following points address their solutions:

- What is the difference between instant segmentation and object detection?

Object detection and semantic segmentation represent the main approaches of object recognition, being able to detect and identify categories of elements in images (Hariharan et al., 2014). Both methods are based on a learning process from imagery, getting to recognize specific features and patterns of objects. The differences between them lie in the resulting outcome. Object detection returns the spatial dimension of the object contained in a bounding-box with the respective label of the item. For its part, semantic segmentation yields a polygon covering exclusively the region occupied by the object itself.

- Which existing model fits adequately to generate the desired segmentation process?

Given that the methodology of this thesis required manipulating the graphical properties of the objects individually, the object recognition model had to be able to extract the objects' borders. The model also needed to be capable of performing the segmentation of instances from a video file, taking into account the image resolution quality of the output file, so that the boundaries of the segmented regions were clearly distinguishable.

Consequently, YOLACT, a recent segmentation-based model that is also capable of producing object detection, was used and calibrated to obtain the object instances. *YOLACT* allowed to satisfy the conditions established to select the object recognition model, in such a way that the instances of the objects in the recording were clearly delimited and labeled. In the same way, the resolution of the output video was high definition, which facilitated the subsequent process of manually masking the segmented regions.

- How to integrate an object recognition model with image enhancement techniques to highlight and downgrade elements in a video?

Since this thesis aimed at working with post-processing techniques, all of them had been integrated so that the workflow could be structured and operational. In the first stage, the calibration of the object recognition model, in this case, *YOLACT*, yielded a video file with instances of several categories of objects that were identified. Subsequently, those detected items were required to be visually emphasized or debased according to their usefulness. Therefore, the second stage necessitated going through each frame making up the video file, and working on certain regions of each image, so that the pixel properties in specific areas could be modified.

The software employed to carry out these tasks was *BLENDER* version *2.83.0*. *BLENDER* allowed breaking up the input into frames, creating 1,818 images from a video lasting one minute. In each frame, it was possible to create masks that recreated the geometry of the segmented objects and traced their movement through the frame sequence. The enclosed regions brought about by the masks were specially treated under the Alpha Over operation, which posed layers on top of others and made it possible to adjust the pixel attributes only in certain areas without affecting the whole scene.

In addition, *BLENDER*, as a graphic design tool, controlled the pixel feature transformation of the segmented regions so that both highlighting and downgrading operations were exclusively applied to the identified objects.

- Can a preprocessed image segmentation tool together with visual enhancement methods facilitate object recognition in *Mixed Reality?*

Through the implementation of two user tests, participants were asked to focus on a virtual overlaid item during the playback of a video filmed in an urban environment. For the development of the first user test, no alterations were made to the video, so that no factors influenced the user's attention. In contrast, the shooting displayed in the second user test was processed in such a way that certain objects were debased by blurring and others highlighted through the change of saturation levels.

After watching the video, users replied to two multi-choice questions regarding what elements drew their attention the most besides the digital overlaid icon. According to the results obtained, the most striking objects remained so even when their visual properties were degraded. On the other hand, the objects that initially turned out to be unattractive caught the users' attention more when they were emphasized.

As displayed in Figures 30 and 31, the users' attention could be reoriented towards elements that needed to be perceived. Therefore, the application of an object recognition model along with visual enhancement methods in *Mixed Reality* environments can improve the conspicuousness of objects while performing an activity, navigation, for example, leading to facilitating its development.

5.4. Drawbacks and recommendations

During the development of this thesis, some technical limitations were encountered. The implementation of the object recognition model produced some misclassifications when detecting certain classes of objects, such as the painting in the background which was recognized as a person. Similarly, the creation of instances from the segmentation process in few cases created geometries that were not identical to the identified object, making the masking process more difficult to accomplish.

Regarding the application of the image debasing technique, due to the visibility of the edges of the objects even though they had already been degraded, it was necessary to increase the deformation of the pixels. This process caused diffuse areas to be seen beyond the segmented regions, blurring part of bordering elements.

Therefore, for future applications, misclassification of objects can be avoided by training the algorithm with more specific object classes. In this way the object recognition model will be able to identify a larger number of elements and more accurately predict the category to which each of them belongs. Also, to improve the segmented regions, some aspects of data collection can be considered, such as video quality, camera movement during recording, and lighting. Considering the out-of-focus areas, the level of blurring can be decreased, so that the visibility of other objects is not hindered.

Likewise, the process of masking the objects in each of the video frames, necessary for the visual modification of the segmented regions, was a time-consuming task. This may represent one of the main constraints to implement this technology in real time. However, for future applications, this process can be integrated into the object recognition model so that instances of objects can be highlighted or downgraded once they are identified.

On the other hand, the user tests also faced limitations. Since participants were asked to select three items from the list in the two questions in each test, in some cases more or less items were chosen. This event forced these answers to be discarded as it was necessary to have a balance between the number of answers in each test in order to compare the change in the users' perception. As a consequence, more participants were required to answer the questionnaires. Additionally, although the tests were not intended to be answered from a particular type of device, some participants reported their inability to play the video recording on their mobile devices.

Finally, the realization of the two user tests took into account the answers of different groups of participants to avoid any kind of bias in the result. However, an approach for a future study involving more than one user test can consider applying them in different temporalities, so that the same participants can estimate and measure the improvement of the methodology that was developed.

6. Conclusions and further developments

Through the use of an object recognition model along with image enhancement techniques, this thesis aimed at facilitating the navigation experience in the context of Mixed Reality by recognizing more efficiently those objects that provide relevant information to users to navigate. The methodology proposed in this thesis was implemented in a video recording in which instances of the objects were detected and extracted. Subsequently, the segmented regions were visually highlighted and downgraded according to their usefulness to fulfill the navigation task.

Likewise, by conducting two user tests, it was possible to evidence the impact on the participants' perception. The results allowed determining that the highlighted objects were more striking and evident to the users. However, a significantly opposite effect did not occur when the characteristics of less relevant objects for the navigation task were de-emphasized.

The contribution of this thesis lies in integrating the field of navigation and image processing from a cartographic perspective. In this way, through the design and modification of instances that represent tangible objects, an abstraction of the real world is created, reducing the users' effort to focus on the elements that need to be perceived.

For further developments in this field of research, some improvements should be considered. The nature of the image segmentation method allows exploiting the information contained in imagery; however, to make the most out of this technology, more and more object classes are required to train the object recognition models. Additionally, more effective techniques for downgrading regions in images need to be explored.

Finally, this thesis opens up the possibility of implementing the methodology developed in real-time models, so that it can be used and integrated with actual applications used for navigation. Additionally, navigation assisted by digital-based content and now supported by object recognition and image enhancement techniques, will make this task even easier for users to accomplish.

7. References

Albawi, S., Mohammed, T., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. *International Conference on Engineering and Technology (ICET)*. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Azuma, R. T. (1997). A survey of augmented reality. *Presence*, *6*(4), 355–385.

Bochkovskiy, A., Wang, C., & Liao, H. M. (2020). *YOLOv4: Optimal Speed and Accuracy of Object Detection*.

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019). YOLACT: Real-time instance segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-Octob*(1), 9156–9165. https://doi.org/10.1109/ICCV.2019.00925

Carmigniani, J., Furht, B., Anisetti, M., Ceravolo, P., Damiani, E., Ivkovic, M., Carmigniani, J., Furht, B., Anisetti, M., Ceravolo, P., & Damiani, E. (2011). Augmented reality technologies, systems and applications. *Multimedia Tools and Applications*, *51*(1), 341–377. https://doi.org/10.1007/s11042-010-0660-6

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 2172–2180. https://arxiv.org/abs/1606.03657.

Çöltekin, A., Griffin, A. L., Slingsby, A., Robinson, A. C., Christophe, S., Rautenbach, V., Chen, M., Pettit, C., & Klippel, A. (2020). Geospatial Information Visualization and Extended Reality Displays. In H. Guo, M. F. Goodchild, & A. Annoni (Eds.), *Manual of Digital Earth* (pp. 229–264). Springer.

Costanza, E., Kunz, A., & Fjeld, M. (2009). Mixed reality: A survey. *Lecture Notes in Computer Science*, 47–68. https://doi.org/10.1007/978-3-642-00437-7_3

Esteban G., M. M. (2012). *Let´s Speed Up! Inglés para Automoción*. Paraninfo.

Fallah, N. (2010). AudioNav: a mixed reality navigation system for individuals who are visually impaired. *ACM SIGACCESS Accessibility and Computing*, *96*, 24–27.

Fast-Berglund, Å., Gong, L., & Li, D. (2018). Testing and validating Extended Reality (xR) technologies in manufacturing ScienceDirect Costing models for capacity optimization in Industry 4.0: Trade-off between used capacity and operational efficiency. *Procedia Manufacturing*, *25*, 31–38. https://doi.org/10.1016/j.promfg.2018.06.054

Franca, M., Kroeker, T., Lambert, N., Malo, E., Michalski, J., Mylnikov, A., Jeffrey, I., & Ashraf, A. (2019). *Smart Object Recognition Architecture A Framework for Supervised Learning Object Detection Group 15*.

Giesemann, F., Payá-Vayá, G., Blume, H., Limmer, M., & Ritter, W. R. (2017). Deep learning for advanced driver assistance systems. In G. Payá-Vayá & H. Blume (Eds.), *Towards a common software/hardware methodology for future advanced driver assistance systems* (pp. 105–132). https://doi.org/10.13052/rp-9788793519138

Gould, S., Gao, T., & Koller, D. (2009). Region-based Segmentation and Object Detection. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (pp. 655–663).

Grasset, R., Mulloni, A., Billinghurst, M., & Schmalstieg, D. (2011). Navigation Techniques in Augmented and Mixed Reality: Crossing the Virtuality Continuum. In B. Fuhrt (Ed.), *Handbook of Augmented Reality* (pp. 379–407). Springer . https://doi.org/10.1007/978-1-4614-0064-6_18

Haralick, R. M., & Shapiro, L. G. (1985). Image segmentation techniques. *Computer Vision, Graphics, & Image Processing*, *29*(1), 100–132. https://doi.org/10.1016/S0734-189X(85)90153-7

Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2014). Simultaneous Detection and Segmentation. *Computer Vision – ECCV*.

Hayder, Z., He, X., & Salzmann, M. (2017). Boundary-aware instance segmentation. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 587–595. https://doi.org/10.1109/CVPR.2017.70

He, Y. (2016). *Object Detection with YOLO on Artwork Dataset*.

Hore, S., Chatterjee, S., Chakraborty, S., & Shaw, R. K. (2018). Analysis of different feature description algorithm in object recognition. In *Computer Vision: Concepts, Methodologies, Tools, and Applications* (pp. 601–635). IGI Global. https://doi.org/10.4018/978-1-5225-5204-8.ch023

Iglovikov, V., & Shvets, A. (2018). *TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation*.

Iwasokun, G., & Akinyokun, C. (2014). Image Enhancement Methods: A Review. *British Journal of Mathematics & Computer Science*, *4*(16), 2251–2277.

Kaur, S. (2013). Contrast Enhancement Techniques for Images-A Visual Analysis. *International Journal of Computer Applications*, *64*(17), 20–25.

Knutsson, P., & Georgsson, O. (2019). *Augmented Reality Navigation Compared to 2D Based Navigation*.

Kunkel, N., & Soechtig, S. (2017). *Mixed reality: Experiences get more intuitive, immersive, and empowering*.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8693 LNCS*(PART 5), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, L., Ouyang, W., Wang, · Xiaogang, Fieguth, P., Chen, · Jie, Liu, · Xinwang, & Pietikäinen, M. (2020). Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision*, *128*, 261–318. https://doi.org/10.1007/s11263-019-01247-4

Liu, Q., & Wu, Y. (2012). Supervised Learning. In *Encyclopedia of the Sciences of Learning* (pp. 3243–3245). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_451

Lorenz, M., Busch, M., Rentzos, L., Tscheligi, M., Klimant, P., & Frohlich, P. (2015). I'm There! the influence of virtual reality and mixed reality environments combined with two different navigation methods on presence. *2015 IEEE Virtual Reality Conference, VR 2015 - Proceedings*, 223–224. https://doi.org/10.1109/VR.2015.7223376

Maini, R., & Aggarwal, H. (2010). A Comprehensive Review of Image Enhancement Techniques. *Journal of Computing*, *2*(3), 8–13.

Mann, S., Furness, T., Yuan, Y., Iorio, J., & Wang, Z. (2018). *All Reality: Virtual, Augmented, Mixed (X), Mediated (X,Y), and Multimediated Reality*.

Medici, C. (2016, January 14). *Virtual reality could become an $80B industry: Goldman*. CNBC.

Mehdi, A., Scuturici, M., Chokri, B. A., & Miguet, S. (2020). A skyline-based approach for mobile augmented reality. *The Visual Computer*, 1–16. https://doi.org/10.1007/s00371-020-01830-8

Miksys, L., Jetley, S., Sapienza, M., Golodetz, S., & Torr, P. H. (2019). *Straight to shapes++: Real-time instance segmentation made more accurate*.

Murphy, C. (2015). Intellectual highlighting of remote sensing imagery for better image map design. *Proceedings of the 27th International Cartographic Conference (ICC'15)*.

Nagata, J., Martínez, A., & García-Bermejo, G. (2017). Realidad aumentada y navegación peatonal móvil con contenidos patrimoniales: Percepción del aprendizaje. *RIED. Revista Iberoamericana de Educación a Distancia*, *20*(2), 93–118. https://doi.org/10.5944/ried.20.2.17602

Nakahara, H., Fujii, T., Yonekawa, H., & Sato, S. (2018). A lightweight YOLOv2: A binarized CNN with a parallel support vector regression for an FPGA. *FPGA 2018 - Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, *2018-Febru*, 31–40. https://doi.org/10.1145/3174243.3174266

Nimkar, S., Shrivastava, S., & Varghese, S. (2013). Contrast enhancement and brightness preservation using multidecomposition histogram equalization. *Signal & Image Processing : An International Journal (SIPIJ)*, *4*(3), 83–92. https://doi.org/10.5121/sipij.2013.4308

O'Shea, K., & Nash, R. (2015). *An Introduction to Convolutional Neural Networks*. http://arxiv.org/abs/1511.08458

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 779–788. https://doi.org/10.1109/CVPR.2016.91

Redmon, J., & Farhadi, A. (2018). *YOLOv3: An Incremental Improvement*. http://arxiv.org/abs/1804.02767

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 6517–6525. https://doi.org/10.1109/CVPR.2017.690

Rokhsaritalemi, S., Sadeghi-Niaraki, A., & Choi, S. M. (2020). A review on mixed reality: Current trends, challenges and prospects. *Applied Sciences*, *10*(636). https://doi.org/10.3390/app10020636

Saleem, A., Beghdadi, A., & Boashash, B. (2012). Image fusion-based contrast enhancement. *EURASIP Journal on Image and Video Processing*, *2012*(1). https://doi.org/10.1186/1687-5281-2012-10

Sandford, A., Sarker, T., & Bernier, T. (2018). Effects of geometric distortions, Gaussian blur, and contrast negation on recognition of familiar faces. *Visual Cognition*, *26*(3), 207–222. https://doi.org/10.1080/13506285.2017.1407853

Shin, W., Kim, M., Park, S., & Baek, N. (2020). A Real-time Video Processing Implementation with Massively Parallel Computation Support. *2020 International Conference on Electronics, Information, and Communication (ICEIC)*, 1–2. https://doi.org/10.1109/ICEIC49074.2020.9051222

Singh, K. K., & Singh, A. (2010). A study of image segmentation algorithms for different types of images. *International Journal of Computer Science Issues*, *7*(5), 414–417.

Smith, S. (1999). *The Scientist and Engineer's Guide to Digital Signal Processing* (Second Edition). California Technical Publishing.

Souza, V. M. A., Rossi, R. G., Batista, G. E. A. P. A., & Rezende, S. O. (2017). Unsupervised active learning techniques for labeling training sets: An experimental evaluation on sequential data. *Intelligent Data Analysis*, *21*(5), 1061–1095. https://doi.org/10.3233/IDA-163075

Speicher, M., Hall, B., & Nebeling, M. (2019, May 2). What is mixed reality? *Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3290605.3300767

Swaminathan, A., Subramaniyan, K. V., Tiruppathirajan, G., & Rajkumar, J. (2020). *Wavelet adaptive quantization based color image segmentation*. TechRxiv. https://doi.org/10.36227/TECHRXIV.12249782.V1

Tacca, M. C. (2011). Commonalities between perception and cognition. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00358

Tadros, M. A., & Franklin, J. B. (2019). *Navigation methods for three dimensional (3D) models in mixed reality (MR) environments. U.S. Patent No. 10,180,735*. Patent and Trademark Office.

Van Hurkman, A. (2011). *The Color Correction Handbook: Professional Techniques for Video and Cinema* . Peachpit Press.

Venugopal, V., Sun, Y., & Brandt, A. (2019). Short-term solar PV forecasting using computer vision: The search for optimal CNN architectures for incorporating sky images and PV generation history. *Journal of Renewable Sustainable Energy*, *11*(6). https://doi.org/10.1063/1.5122796

Yang, M.-H. (2009). Object Recognition. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems*. Springer. http://faculty.ucmerced.edu/mhyang

Yoon, H., Han, Y., & Hahn, H. (2009). Image Contrast Enhancement based Sub-histogram Equalization Technique without Over-equalization Noise . *International Journal of Electrical and Electronics Engineering*, *3*(6), 323–329.

Zhang, Y. (2017). *Image Engineering*. De Gruyter.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 6230–6239. https://doi.org/10.1109/CVPR.2017.660

Zhao, Z.-Q., Zheng, P., Xu, S.-T., & Wu, X. (2019). Object Detection with Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(11), 3212–3232.

Zhu, Z., Li, D., Hu, Y., Li, J., Liu, D., & Li, J. (2020). Indoor scene segmentation algorithm based on full convolutional neural network. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-020-04961-0