3D Estimation of Visual Focus of Attention

Carlos Miguel Antunes Simões
carlos.miguel@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

November 2021

Abstract

Humanoid and social robots may provide valuable resources to society in the most diverse and complex activities and challenges, thanks to their increasing mechanical and decision-making abilities. However, robots must comprehend and acquire information about their surroundings for proper interaction with humans. The VFOA can be used as the primary conversational cue. To tackle these challenges, we develop a novel approach that estimates and tracks the VFOA. The proposed model stems from the consideration that the eye gaze and head pose carry information about actions and interactions. The proposed formulation leads to a 3D algorithm that considers: (i) A bounding box of every object in the field of view of the robot's camera, and (ii) a ray casting algorithm that considers head and gaze directions. A Kalman filter performs the tracking of the gaze. Finally, the VFOA algorithm estimates the object of attention based on a weighted sum of gaze and head pose information. We study the parameters of 3D VFOA algorithm, running simulated scenarios for a selection of the most adequate parameters. The novel approach is validated, tested and benchmarked on the public MPI Sintel dataset containing animated real-world interactions.
Keywords: VFOA, Eye Gaze, Head Pose, Object Detection, Human-Robot Interaction.

1. Introduction

Currently, robots thrive in rigidly constrained environments, such as factory plants, where human-robot interaction is kept to a bare minimum. A long-term goal of researchers in robotic systems is to aid in the transition of this field into our daily lives, namely to our households. Ultimately, we strive to achieve cooperative, effective, and meaningful interactions. Usually, two or more individuals try to communicate with each other by exchanging messages during an interaction, which are typically associated with speech. However, while verbal communication may be the most obvious form of communication among humans, non-verbal cues such as body posture, gestures, facial expressions, intonation, gaze direction, and more often convey considerable amounts of information, as stated by Breazeal [2].

This thesis's main objective is to estimate the VFOA, i.e., identifying both the perceiver and their visual target during an interaction, which is recognized as one of the most notable social cues, stated by Bernard Ogden [11]. Knowing a person's VFOA can be used in social interactions to establish face-to-face dialogues, prevent speaking over someone, or to draw someone's attention, which can help to initialize or maintain fluid interactions.

Therefore, determining the gaze or the head pose (or both) is necessary to identify the visual target. This implies the estimation of an imaginary line emitted from the perceiver's face to their target. In other words, the direction of the eye gaze or its head. According to Stiefelhagen et al. [13], eye gaze alone is insufficient to calculate the focus of attention of an individual. However, accurate results may be determined when head positioning is also taken into account. Accordingly, in this thesis, we are interested in estimating the VFOA based on the head pose and gaze in a scene, to allow a robust and efficient interaction. Some of the benefited fields are **Psychology and Sociology** where the VFOA is often used to guess the object of attention, which can be utilized later to begin a conversation. **Human-Robot Interaction** a robot estimates and understands the VFOA to enable smooth and pleasant interactions between the robots and persons. An example can be the robot explaining a piece of art to a person that was focus on.**Virtual Reality**, **Assisted Systems** or **Web Interface Design** can also benefit for gaming, autonomous driving or simply the position or rearrangement of publicity/highlight content.

The remainder of the document is structured as follows: Chapter 2 presents a brief background about previous work. Introducing the current state of the art for object detection, gaze and head pose

estimation, and finally, attention.

Chapter 3. In this chapter, we introduce the framework used to estimate the VFOA from the gaze and head pose estimation. Since the implementation steps to obtain the 3d object from the stereo acquisition, its reconstruction passing through the image segmentation, to the novel approach created to obtain the VFOA.

Chapter 4 ultimately discusses the synthetic experiments and presents the results of the MPI Sintel dataset on the achieved VFOA.

Finally, Chapter 5, we discuss and analyze the main contributions produced and conclude the work of our findings. Finally, we present suggestions for potential future work.

2. Background and Related Work

2.1. Object Detection State of Art

Object Detection general purpose is the identification of objects of a specific class (such as people, animals, or vehicles) as well as their location in images and videos. Convolutional neural networks have been commonly used as the backbone and detection network of most state-of-the-art object detectors to derive features from input images and/or recordings, classification, and localization, respectively.

2.1.1 Classification

Classification consists in identifying the class that the corresponding object belongs. LeNet was probably the first CNN approaches introduced in 1989, with the objective of digit classification. Then through competitions, the models continue to improve their results and increase the number of layers and weights. The first significant improvement appears when ResNet introduces the 'shortcut' module surpassing human-level accuracy.

2.1.2 Detection

Object detection consists of identifying the object's class along with its respective location in an image, usually through a bounding box. For this task, more challenging and demanding models were created such as R-CNN [5], and Fast R-CNN [4]. The novelty was the use of selective search to obtain the region proposals.

2.1.3 Segmentation

Image segmentation is a technique utilized to divide an image into several areas based on pixel characteristics providing a pixel-wise mask for each object in the image. Therefore models such as Faster R-CNN [12], and Mask R-CNN [6] were introduced, outperforming all the other single-model state of art methods these tasks.

2.2. Head Pose State of Art

The head pose can be formally defined by Euler angles, in the Tait-Brayan notation, these three angles can be called pitch, yaw, and roll, which refer to the rotations along the x,y, and z-axes, respectively. As a result, the whole purpose of the head pose estimation is to extract these angles from the head or face that was being observed. Head Pose estimation is a hotly debated subject since the majority of researchers agree that extracting the head pose is simpler than estimating eye gaze and more likely to be seen in real-world situations. Is commonly acknowledged that head pose techniques have been classified into several groups. For example, a classification-based method that associates each image with a discrete head pose label and a regression-based approach that learns how to map a head pose image to a head pose result (or by simply deriving from facial features).

2.3. Gaze State of Art

The gaze estimation issue can be divided into three categories: gaze fixation point estimation, which determines the human gaze's fixation point in 2D on a specific flat surface. Gaze following, which tries to deduct the objects people are staring at, and finally, 3D gaze estimation, which estimates the gaze positions on a screen or an image.

2.4. Attention State of Art

Estimating the VFOA is identifying what are we looking at or who is looking at whom, which can be recognized in a multi-party dialog or in within any scenario as one of the most prominent social cues. Therefore to calculate the VFOA several methods were proposed, some using only the head orientation, others tries to use both the head pose and gaze together. However the estimation of the gaze can be achieve directly through the image of the eyes or through algorithms that uses the head pose to discover the gaze direction.

3. Implementation

3.1. Architecture Overview

The main goal of this work is with the extracted information obtained from both gaze and the head poses estimate the VFOA, allowing the initiation of fluid interaction.

As stated in section 1, we define the VFOA as identifying the visual target, e.g., person or object, on which the person of interest is focused. For this purpose, it is essential to detect the person's, and more specifically, the headbox and objects. The headbox allows us to estimate the head pose, represented by a 3D vector containing pan, tilt, and roll angle, and the person's eye direction, known as gaze. In this project, we present an estimation for the VFOA based on gaze and the head pose orien-

tation.

The proposed formulation leads to a 3D algorithm inspired by ray casting that reconstructs the 3D environment and then estimates the VFOA. Therefore a reconstruction procedure is necessary to build the 3D environment based on detected objects. It begins with the stereo vision system that enables us to acquire the disparity map. Then, a panoptic segmentation model detects and segments all the objects in the image. Combining this information with additional transformations enables us to reconstruct the scene to a 3D map.

The Architecture overview of the complete framework is displayed in the Figure 7.

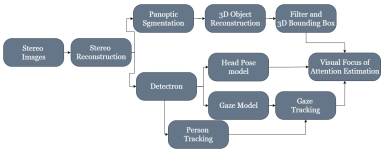

Figure 1: Thesis architecture diagram.

3.2. Stereo Reconstruction

The stereo reconstruction corresponds to the extraction and calibration of the information given by a pair of images, known as stereo images, provided by a pin-hole camera. A pipeline is created, starting with the camera calibration phase, where the estimation of the intrinsic and extrinsic and the lens distortion parameters are acquired. These two images are then inputted to a rectification algorithm, which then removes the lens distortion effects. The second step uses both images to find the corresponding pixels between these two views by creating a so-called disparity map. The algorithms considered in this thesis were StereoSGBM and StereoBM. An intermediate step corresponds to a filtering step for both left and right images essential to enhance their qualities and filter out unwanted image noise. The filter selected for this task was the WLS. The third step in this pipeline uses the filtered disparity map to triangulate the left and right images and find the 3D coordinates. It's important to refer that this triangulation requires the knowledge of the stereo camera calibration parameters, which were achieved in the first step. The current implementation was based on Bradski's book [1]. Finally, the last step consisted of visualizing the reconstructed point cloud, this was generated using the Open3D Library [17].

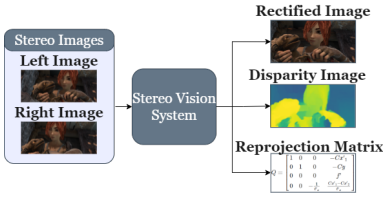

Figure 2: The Stereo Vision system.

3.3. Image Segmentation

Identifying all potential objects is a critical phase in this project to determine the actual object of attention. As a result, an image segmentation method is required.

Image segmentation can be divided into two mains parts: semantic segmentation, which represents the method of associating each pixel of an image with a class label, and instance segmentation, which, unlike semantic segmentation, masks each instance of an object in an image. A novel approach in image segmentation is Panoptic segmentation, which is the combined version of two primary segmentation methods. Panoptic segmentation was the method chosen because, in a single task, it provides the location in the image of the object and its classification, potentially identifying and locating the object of focus for effective interaction in one model.

The model utilized in this project was made by Kirillov et al. [9], which provides two channels: one for pixel's label, which represents the semantic segmentation, and another for predicting each pixel instance, resulting in instance segmentation) that are going to be used in the 3D Reconstruction and Attention model.

3.4. 3D Object Reconstruction

The final reconstruction step is the reconstruction of the objects identified by the Panoptic Segmentation model. Therefore, instead of performing reconstruction on the entire image, we perform only on the ROI detected. This approach has two key advantages: 1) reconstructing only the ROI provides several regions that will not be reconstructed, which reduces the noise in the reconstruction, improving runtime speeds; 2) by performing reconstruction on ROI, we reduce the possible range of unidentified objects of focus. As a result, the interaction can proceed more smoothly and facilitate identifying the object of focus in the Attention algorithm. Joining the instance segmentation mask for each detected object in the rectified image and the respective classification generated from the panoptic segmentation, together with the stereo matching output, results in several instance disparity maps containing only the objects of interest. Finally, the instance disparity map is reconstructed into a point cloud. Lastly, with the instance point cloud, we can

filter and estimate the 3D bounding boxes for each object/person detected.

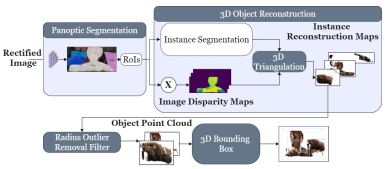

Figure 3: 3D Object Focus Reconstruction Process.

3.5. Gaze Estimation

Several datasets were taken into account to obtain the most reliable model for gaze estimation in 3D. Analyzing the performance of each model, it is possible to conclude that the model provided by the Gaze360 presents the best results to every dataset, and because of that, in this project, the model Gaze360 [8] was chosen to estimate the 3D direction of the gaze. One other reason behind this selection was the ability of solving one of the major problems in the majority of the gaze models, that is the ability of dealing with unconstrained scenarios, the partial occlusion of the eye, medium distances, outdoor environments, head and gaze variations. However, the open-source code for the Gaze360 [8] was deprecated and not functional. Consequently, the rewrite and upgrade of the code were done. The implementation consisted in performing the process necessary to extract the detectron2 [15] output that enabled the creation of the headbox, which contains the coordinates and image of the head, a step essential to provide for the gaze model. The tracking of the persons was additionally implemented to take advantage of the temporal model in the Gaze360 and the gaze tracking method implemented.

3.6. Person Tracking

The implemented approach to do the tracking of the persons was the IOU Tracker. The IOU tracker associates person detections of consecutive frames based on their spatial overlap to do the tracking. The proposed approach can be considered greedy: The first frame is created the initial track, then the subsequent detection is associated with the track with the highest IOU or is superior to a certain threshold.

3.7. Gaze Tracking

The implemented method for the gaze tracking was based on the work published by Toivanen [14]. However, some changes were performed. In the Toivanen [14] method, the estimated velocity was derived from an entire picture of the eye as input. In a real-world application, a detailed image of the eyes with quality enough for this type of input is impracticable. So in this thesis, the gaze velocity is calculated utilizing the derivative of the position. The formulation of the measurement noise matrix (R^t) was also adjusted because it also depends on the velocity.

3.8. Head Pose Estimation

The model chosen for the estimation of the Head Pose task was the FSANet [16] since the method results outperformed the state-of-the-art methods studied for both the landmark-free ones and landmark-based or depth estimation. Furthermore, the model only required a single RGB frame as input, and it states that the memory overhead was 100 times smaller than the former methods.

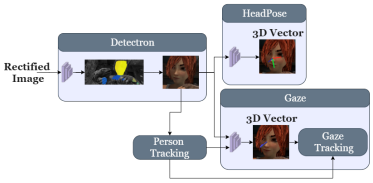

Figure 4: Gaze360 and Head Pose architecture.

3.9. VFOA Estimation

Our objective is to calculate and monitor the VFOA of the people present in a given environment, relying on eyes gaze, and head pose. Thus, we assume to have a specific set of visual targets, provided by the image segmentation/panoptic model, which are of interest in the given scenario, and would like to identify which targets a person or multiples persons are looking at. However, if the image segmentation model excludes some object in the scene, it is not considered.

Identifying the person and its focus target in a general context provides valuable information about their Attention to the receptor robot or a person. The model provided the possible object of focus that could be used to judge how to proceed in an initial conversation.

Eyes 3D Location We propose an Attention algorithm based on ray casting. Therefore, an essential intermediate step is to estimate a valid point for the origin of the rays. The origin point should be the same for the gaze and head pose and should be located between both eyes. To estimate the location, we get the map of the headbox with its reconstructed coordinates in 3D and extract a square map located in the middle horizontal and 65% on the vertical of the headbox providing the estimation for the eyes map. Having the estimated eyes map, we do the median on the 3D reconstructed coordinates to estimate the origin for the rays.

Field of View Construction In the reconstructed environments, where the 3D coordinates of objects are maintained, we choose a 3D mesh (nominally a cone) to represent the field of focus. A person's region visible through his or her eyes

is referred to as the field of view. As Koslicki et al. [10] describe, the field of view can be divided into several parts. The essential region to this project is the field of focus, which represents the person's amplitude of focus and solves the uncertainty in the gaze position. With this knowledge, it is possible to define the amplitude of the cone having the certainly we are considering only the region of focus. The scaled cone is constructed with its vertex at the middle of the eyes of the identified person, and its axis is aligned with the gaze and head pose direction in the world coordinate system as obtained from the gaze and head pose model. By projecting the cone centered on the estimated gaze/head pose direction, it is possible to produce a list of Object-Of-Interest that the observer could be looking at. In order to identify all objects in the region around the gaze/head pose vector, the cone has to be renewed every frame based on gaze direction. The methodology investigated and implemented is based on the ray casting idea, which emits several rays only on the boundary around the estimated gaze/ head pose direction with a radius adjusted to the field of focus (creating a 'hollow' cone).

Intersection Ray Bounding Box Algorithm The slab method was the approach chosen to calculate the intersection of the Bounding Box and the ray. The idea behind the method is to consider the box as a space inside of three pairs of parallel planes. Furthermore, each pair of parallel planes corresponding to the box's margins cut the ray, and if a part of the ray persists, that specific ray intersects the box.

Attention Estimation A novel approach is presented to estimate the attention values, using a collider object algorithm to detect the object of interest and calculate the person's attention to the specific target. With the collision of a ray with a bounding box, we automatically obtain the name of the possible object of attention and the point of intersection. Taking the point of intersection, we calculate the normalized distance to the centroid of the bounding box. However, the points closer to the centroid would be represented with a lower number of Attention. Therefore, we subtract the normalized distance by 1.5, which reverses the importance. This distance provides relative attention to the object. The main idea is that the further away the intersection point is from the centroid less attention is assigned to the object.

$$A_{IG} = \begin{cases} 1,5 - \frac{\|P_I - C_{BB}\|^2}{|P_I - C_{BB}|}\omega_G & \text{, if intersects} \\ 0 & \text{, if no intersection} \end{cases} \quad (1)$$

$$A_{IHP} = \begin{cases} 1,5 - \frac{\|P_I - C_{BB}\|^2}{|P_I - C_{BB}|}\omega_{HP} & \text{, if intersects} \\ 0 & \text{, if no intersection} \end{cases} \quad (2)$$

After checking all the rays to the bounding boxes, we obtain an array for the head pose and the gaze with the distance values for each bounding box. Each array is now summed, providing the total attention for a specific bounding box/ object and the gaze and head pose.

$$A_T = \frac{\sum_{k=0}^{n_R} A_{I_{Gk}}}{n_R\left(1,5 - \frac{d_{min}}{d_{max}}\right)\omega_G} + \frac{\sum_{k=0}^{n_R} A_{I_{HPk}}}{n_R\left(1,5 - \frac{d_{min}}{d_{max}}\right)\omega_{HP}} \quad (3)$$

All the rays emitted from the eyes location, within the field of view have the same relative weight. The only weights difference arises when the representing cones describe the gaze or head pose field of focus. These weights are based on Roxane et al. [7] work, which studies and validates through experiments the relation between the gaze and head pose in two situations, 1) when the angle between the gaze and head pose is lower than 30° and 2) the opposite when the angle is higher than 30°. The conclusion arrived for the first case; the weights given were 60% to the gaze and 40% for the head pose, meaning equal importance to both objects detected through the gaze and the head pose. For the other case, the weights given were 90% to the gaze and 10% for the head pose, meaning that the eye region provides the preponderance influence in detecting the object of focus. Therefore, after calculating the total attention for the bounding boxes, the weights are applied to the rays having the angle between the gaze and the head pose decide the weights applied.

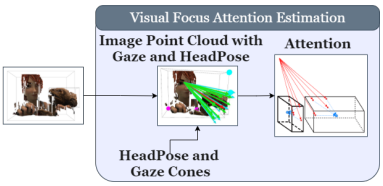

Figure 5: VFOA architecture.

4. Model Validation and Experimental Results

4.1. VFOA Model Validation

Experiments with simulated data were performed to validate the implemented attention block. Two types of experiment were performed: the first type evaluates a *single-object focus activity*, wherein a simulated subject fixates gaze and head pose on a single, fixed point, at different positions throughout an object's bounding box; the second type of experiment evaluates *the interaction involving a subject*

and several objects to assess the application of the methodology to a more realistic task wherein multiple targets are present in a given scene.

The objective of this experiment was to assess the reliability of the estimation of a subject's VFOA, given 1) an estimated orientation of head and 2) an estimated 3D vector of gaze. The goal is to recreate naturally occurring scenarios, e.g., having a subject look directly at a specific target placed in scene and varying the orientation of their gaze and the head pose to compare and evaluate the computed attention with the known object of interest.

Each experiment comprises multiple scenarios with unique characteristics, namely 1) the direction of the gaze, 2) the direction of the head pose, and 3) the intrinsic parameters (number of emitted rays, radius) of the created cone.

Single Object Experiment - Case 1

The first case represents a realistic situation where the subject is looking directly at one object of interest. More specifically, the orientations of the gaze and the head pose overlap in the center of the object.

We observe that the number of rays and the cone radius are essential for quantifying the value of the attention. We conclude that a greater number of rays results in increased sensitivity in the experiments. In contrast, the inverse is observed in relation to the cone's radius: an increase in radius results in a decrease in the attention values, as a result of a reduction in the overall importance of both the gaze and the head pose components.

In conclusion, the best accuracy values for the model are obtained when the radius of the attention cone is between 5 and 30 (i.e., low to average values). Since the rays for the head pose and gaze have the same direction the values of attention should have been the similar, however the disparity between the attention values for the gaze and the head pose is evident due to the relative importance given to the two by the model when they overlap (60% to the gaze and 40% for the head pose mentioned in Attention Estimation).

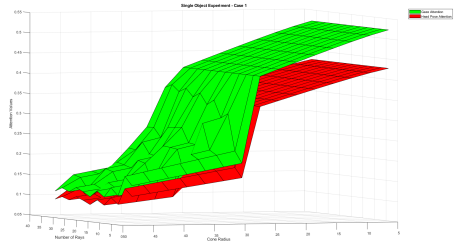

Figure 6: Results for the single object focus scenario, where the subject looks directly at one object, varying the cone's radius and number of rays.

Single Object Experiment - Case 2

In this scenario, the gaze and head pose continue to be directed at the object, but they no longer overlap; instead, the gaze vector is oriented at the center of the object, whereas the head pose is focused on an off-center point at the corner of the object's bounding box. This corresponds to a situation where the subject's gaze and head pose do not match, but the subject is still looking at the object. In addition, the reverse scenario is also considered for this second case. In this scenario, it is the head pose that is focused in the center of the bounding box, whereas the gaze is directed at the corner of the object.

The Attention values obtain for this scenario are similar to that of Case 1, except when the values of the total attention for the head pose, those are lower than in the first case. This observation is explained by the head pose being oriented at the corner of the object, which leads to a decrease in the level of attention relative to the first simulation. This also validates the model's calculation of attention, since the farther the vector of focus is from the centroid of the object, the less influence results in the output attention variable.

In the scenario where the head pose is directed at the center of the object, the attention value for the head pose is considerably higher than the attention given by the gaze, considering that all the rays continue to remain to intersect the object for the head pose vector as well as the gaze vector, and given the variation in the weights given initially for the attention when the rays have more than 30° in amplitude this difference can be explained. For higher values of the cone's radius, the attention for the gaze begins to have more importance than the head pose, since the rays are now closer to the object's center. In addition, the head pose rays begin to get farther away from the centroid of the object. However, the difference between the two is very low because the head pose has more rays on the object than the gaze.

Single Object Experiment - Case 3 Case 3 models the situation where a subject's focus of attention cannot be directly determined, i.e., the gaze and the head pose are directed in opposite directions, one outside the box and the other focus on the object.

We draw two key observations. First, the observation of only a single type of attention value (i.e., gaze or head pose), as shown in the plot, is expected since only one attention cone (either the gaze or the head pose) is attending to the object. The total attention is therefore equal to the attention of that individual component. Second, the attention values begin to decrease as the cone radius increases, due to the expansion of the intersection rays and

the subsequent increase in the distance to the centroid. This experiment also models the case where the system only encounters one object in the image and the gaze is not directly focused on that object; however, if the head pose is oriented at it, then the VFOA is determined entirely by the head pose.

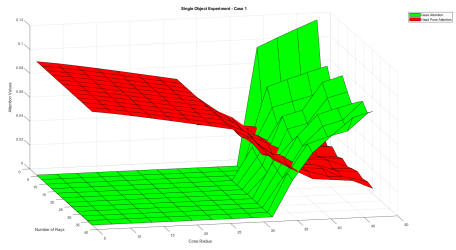

Figure 7: Results for the simulated experiment for the single object focus where the subject has their head pose placed on one object, but their gaze oriented outside of the bounding box, with varying cone radiuses and number of rays.

Multi Object Experiment - Case 4 Each of the previous experiments analysed a single object. However, in a real-world environment one expects to find more than one object. The following experiments (Cases 4-5) model such cases.

The first case explores the case where, as in the first experiment, a subject simultaneously looks at two objects. Low to average cone radiuses yield better precision in determining the subject's attention, and the influence of the number of rays is important to disambiguate (if necessary) which object is primarily focused. For the current experiment, the values of attention are similar between objects. However, the system tends to give a slight advantage to the bigger object, since the rays spend a greater proportion of time on the object.

Multi Object Experiment - Case 5

This experiment models a scenario where the head pose and the gaze focus on different objects. This situation, occurs several times in real-world interactions, usually when a subject tries not to openly reveal the real object they are looking at.

This experiment shows that even in complex scenarios such as when a subject tries to hide their focus on an object of interest by orienting their head in a different direction, the model distributes the attention to both objects while giving greater importance to the object the gaze is focused on.

VFOA Parameters Tuning

The experiment scenarios described in Section 4.1 try to represent all possible real-world interaction scenarios. It is possible to derive *quantitative* parameters which can be used to assess and fine-tune the attention model's performance. These parameters are determined by summing all the previous Total Attention values across all scenarios. It is then possible to find a global parametrization that describes a meaningful value for all interaction scenarios. These attention values can be used to derive a *qualitative* analysis of the Attention model's accuracy as a function of its tuning parameters.

Several parameters can be tuned to improve the object detection accuracy and their corresponding attention values, as determined by the algorithm that determines the VFOA. The experiments evaluated two parameters: 1) the *Number of Rays* emitted at the origin and directed at either the gaze or the head pose, and 2) the *Field of View*, which determines the rays' geometry. In the experiments, the Number of Rays was initialized with a value of 4 and increased until 40. Finally, the Field of View was initialized with a value of 5.2° (corresponding to a cone radius of $5cm$) and increased until 46.3° ($47cm$).

We observe that increasing the Number of Rays improves the overall accuracy of the estimation procedure. In particular, if the circumference of the cone's base does not contain a sufficiently high density of projected rays, there will be several spaces that will not considered, effectively creating blindspots; this results in some objects being identified incorrectly or not at all. Therefore, increasing the number of projected rays yields a greater chance of identifying the objects accurately. In contrast, an increase in the Field of View impacts the value of attention negatively. This is consistent with the observations of Koslicki et al. [10], who identified the different human fields of view, including the *field of focus*, which corresponds to a 30° field of view angle. This experiment demonstrates that lower values corresponding to the field of focus (i.e., roughly between 5° and 25°) yield the optimal attention values for the algorithm.

4.2. VFOA Experimental Results

Experimental Dataset. The model implemented in this thesis was evaluated on a 3D-animated short film from the **MPI Sintel [3]** dataset, on the clip "Bandage" of stereo images, where the interaction is performed by two main characters: a girl and a bird-like dragon. In this set of images is possible to manually annotate the subject's VFOA. The dataset also provides intrinsic and extrinsic camera parameters. One 50-frame clip from the stereo image dataset closely approximates a hypothetical real-world situation. Figure 8 shows the visualization component, with subjects and object identified from the stereo images and reconstructed in a 3D view. The head pose and gaze cones are included in the visualization to illustrate the calculation of attention by representing the subject's field of view and the intersection between the rays and

the bounding box.

Accuracy Measurement. Accuracy was measured using FRR, which corresponds to the percentage of frames for which the VFOA is identified accurately, i.e., matches the ground truth label.

Experimental Overview. The following experiments demonstrate the crucial role played by the stereo component and the number of cones for the determination of the VFOA. For both the single-cone and multi-cone scenarios, both the BM and SGBM stereo matching algorithms were evaluated. Overall, the BM stereo matching algorithm performs faster, whereas the SGBM algorithm was consistently more accurate. In the previous validation results, the cone radius was identified as a critical feature, and we observed that small to medium radii result in the best performance. Therefore, the single-cone experiments consider one cone with small field of view for the gaze and for the head pose; in contrast, the multi-cone experiments consider two cones (one with small field of view, the other with medium field of view) for the gaze and for the head pose. The motivation behind the experiment with two cones is to check if the smaller object go undetected and if the restriction of the field of view influence the Attention estimated.

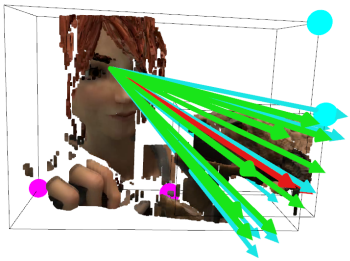

Figure 8: MPI Sintel video "Bandage" reconstructed in 3D view with 1) the subject and the bird identified by bounding boxes, 2) the cone for the gaze (in green), and 3) the head pose (in blue). Both the gaze and the head pose were used to calculate the VFOA.

MPI Sintel Experiment With Single Cone

The following experiments evaluate the scenarios using the single-cone component by considering either the BM or the SGBM stereo matching algorithms.

Stereo BM algorithm.

Table 1 presents a summary of the results provided by the Attention model for each frame. Frames in which the panoptic segmentation model did not recognize any objects are not shown. Attention values corresponding to 0 indicate objects that were not attended to, i.e., objects not involved in the interaction. The last case, where objects different from the bird were identified, represents the incorrect detection of an inexistent object in the frame.

With one cone and using the BM matching algorithm, 28 frames are classified correctly, and 22 frames are classified incorrectly, yielding a total accuracy of 56%. The results show that the head pose was more accurate in finding the object of attention than the gaze, contrary to the weights given to the model based on each model's state of the art. However, it is imperative to note that the use of animated images can negatively influence the models' accuracy, since both the gaze and the head pose models were trained using real-world video recordings. This is one possible reason for the comparatively low Total Attention values shown for most frames. Two key observations can be made by examining the results: first, the head pose is the predominant component for the determination of the Total Attention, since the gaze component is often zero; second, in the ideal situation (i.e., when both the gaze and the head pose are correctly detected, as happens for instance in frames 37 and 38, for which the Total Attention is 0.69 and 0.83, respectively), the total attention values are much higher, and the model is able to correctly determine the Total Attention corresponding to the correct objects with much more robustness.

Stereo SGBM algorithm.

This case preserves one cone; however, the more accurate SGBM matching algorithm is used to discover the model's relation with the 3D reconstruction provided by the disparity. In the final output, the object of attention was correctly identified in 35 of the 50 frames and incorrectly identified in 15 frames, yielding a 70% attention classification accuracy. A 14% improvement was observed by employing a more accurate stereo model. However, the head pose continues to have a more prominent influence than the gaze on the final attention results. This confirms that the gaze model provides a sub-optimal output for the eye orientation in this video.

MPI Sintel Experiment with Multi-Cones

The parameter cone radius in the validation experiments was a critical feature when experimenting on real-world scenarios, primarily due to representing the field of view. This experiment demonstrates the importance of this parameter. In the previous experiment, the attention algorithm only used one cone to calculate the value of attention. However, this analysis increased the number of cones to two, and the radius varies from small to medium. These values were chosen based on the analysis in the evaluation section. Both radius values showed the best results and matched the definition of the focus field of view.

Figure 9 displays the two cones and the char-

acters' interaction in the MPI Sintel "Bandage" dataset.

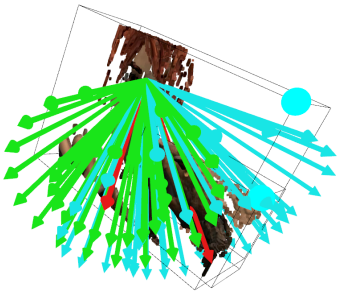

Figure 9: MPI sintel "Bandage" video reconstructed in 3D view with the person and the bird correctly identified by the bounding boxes, and the two cones each, for the gaze, green, and the head pose, blue. The Rays of the cones were used to calculate the VFOA.

Stereo BM algorithm.

For the case of the BM matching algorithm, the results correspond to a 62% classification accuracy (corresponding to a 6% improvement from the previous case) associated with the correct identification of 31 frames (instead of 28) while using the same matching algorithm. This result demonstrates that it is essential to have as many cones as possible within the focus field of view. This conclusion can be explained easily by observing that a greater number of rays within the field of view increases the probability of finding the objects through the attention model's ray casting algorithm.

Stereo SGBM algorithm.

This last case preserves the two cones; however, the matching algorithm was again switched to the SGBM algorithm. Therefore, this experiment employs the optimal approaches considered in this project for depth estimation: the SGBM algorithm coupled with the optimal parameters obtained for the attention model, the cone radius, and the number of cones. The object of attention was correctly determined for a total of 38 frames, corresponding to a final classification accuracy of 76%. The use of the superior matching algorithm provides the expected accuracy improvement of 14% compared to the MC-BM case. The attention values keep the same average, which confirms the proper normalization of the attention model. Compared to the same stereo model with a different number of cones, the Frame-based Recognition Rate increases 6%. This represents a constant improvement afforded by the use of multiple cones to accurately predict the object of focus. It is important to note that six video frames provided no Attention values or objects over all the experiments because the segmentation model incorrectly classifies the scene as having no characters. Therefore, the model is not able to reconstruct the interaction correctly, i.e., it is not able to recognize the correct object of focus (the bird in the scene), nor is it able to calculate adequate Attention values.

Number of Cones	Stereo Algorithm	# of frames with VFOA Correct	# of frames with VFOA Wrong
Single-Cone	BM	28	22
	SGBM	35	15
Multi-Cone	BM	31	19
	SGBM	38	12

Table 1: Summary of the VFOA Experimental Results.

5. Conclusions

This dissertation has strived to determine the VFOA based on the orientation of the eyes and the head. Specifically, we sought to disentangle the role of gaze direction and head orientation by adjusting their relative weights in attention estimation depending on a given situation. Furthermore, several synthetic experiments were executed to validate the model, by presenting a simulation of possible real-world situations, for the gaze and head orientations. Also, through the experiments, we present a novel approach to calculate the VFOA over the 3D reconstruction of a scene through stereo images. During those experiments, we also show that the parameters, such as the cone radius (i.e., the field of view) and number of rays have a critical influence on the Attention estimation results. Greatly due to the rays approach on the 3D reconstruction of a scene.

Moreover, the algorithm verifies the definition of the *field of focus* on a real-world simulation. The *field of focus*' amplitude was, through the experiments, proven as a crucial parameter for detecting the correct *object of focus*. Taking only the vector that provides the direction of the eyes' gaze and head pose into account can deliver the incorrect *object of focus* or the incorrect level of attention. In conclusion, emitting as many rays as possible within the amplitude of the field of focus can increase the final values of attention because it reflects the totality of a person's field of focus, capturing all the possible objects of interest.

Reconstructing the scene in the 3D world can be helpful to interpret the results, by helping explanation and validation of the object taken as the focus. Another significant advantage lies in accurately and dynamically detecting the object of focus because, by reconstructing the scene in a 3D world, we get a more realistic environmental perspective. Additionally, our model's attention values can be distributed throughout the objects on the scene, originating different weights that provide a more accurate evaluation of the object of focus and its evolution over

time. Therefore the correlation between the accuracy of the reconstruction and the attention model's is high, so, if the accuracy of the reconstruction increases, the attention model tends to present more accurate levels of attention.

Finally, the experiments on the images in the MPI Sintel dataset shows promising results of this novel approach for calculating the attention's visual focus. The models applied for image segmentation, gaze, and head pose were trained on real-world people and environments. However, the images in this experiment's dataset, provide an animated scene in which, although the models provided less accurate information to the attention model, it was still robust enough to detect the object of focus accurately in almost 80% of the frames. The attention model provides a level of attention for each frame independently of the size of the bounding box, providing a normalised attention value, resulting in comparable levels of attention between bigger and smaller objects. A significant drawback of the approach is not providing the results in real-time. Therefore a future investigation of transforming the developed methods to a real-time interaction would be a priority.

References

[1] G. Bradski and A. Kaehler. *Learning OpenCV: Computer vision with the OpenCV library.* " O'Reilly Media, Inc.", 2008.

[2] C. Breazeal. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59:119–155, 07 2003.

[3] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, pages 611–625, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[4] R. Girshick. Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 International Conference on Computer Vision, ICCV 2015:1440–1448, 2015.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[6] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:2980–2988, 2017.

[7] R. J. Itier, C. Villate, and J. D. Ryan. Eyes always attract attention but gaze orienting is task-dependent: Evidence from eye movement monitoring. *Neuropsychologia*, 45(5):1019–1028, 2007.

[8] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. 2019.

[9] A. Kirillov, R. B. Girshick, K. He, and P. Dollár. Panoptic feature pyramid networks. *CoRR*, abs/1901.02446, 2019.

[10] W. Koslicki, S. Babin, D. Makin, R. Vogel, J. Contestabile, and K. Kohri. Resilient communications project: Body worn camera perception study phase 1 memorandum report, 07 2018.

[11] B. Ogden and K. Dautenhahn. Robotic etiquette: Structured interaction in humans and robots. *Proc. SIRS2000, Symposium on Intelligent Robotic Systems, Reading, UK*, (1998):353–361, 2000.

[12] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[13] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel. From gaze to focus of attention. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1614:761–768, 1999.

[14] M. Toivanen. An advanced kalman filter for gaze tracking signal. *Biomedical Signal Processing and Control*, 25:150–158, 03 2016.

[15] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[16] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] Q.-Y. Zhou, J. Park, and V. Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.