2022 IEEE International Conference on
Autonomous Robot Systems and Competitions (ICARSC)
Santa Maria da Feira, Portugal – April 29-30, 2022

3D Estimation of Visual Focus of Attention

Carlos Miguel Antunes Simões
Instituto Superior Técnico
Lisboa, Portugal
carlos.miguel@tecnico.ulisboa.pt

Plinio Moreno
Instituto de Sistemas e Robótica
Instituto Superior Técnico
Lisboa, Portugal
plinio@isr.tecnico.ulisboa.pt

***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.**

***Index Terms*—VFOA, Eye Gaze, Head Pose, Object Detection, Human-Robot Interaction**

I. 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 [1].

This work presents an approach to estimate the Visual Focus of Attention (VFOA), i.e., identifying both the perceiver and their visual target during an interaction [2]. Estimation of VFOA through gaze and head pose, will allow social robots 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. An example of such an interaction is a robot that explains a piece of art to a person that was focus on.

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. According to Stiefelhagen et al. [3], eye gaze alone is insufficient to calculate the focus of attention of an individual. However, accurate results may be determined when head pose is also taken into account. Thus, we are interested in estimating the VFOA based on the head pose and gaze in a scene, to allow a robust and efficient interaction. The VFOA in this work is estimated through a weighted combination of attention from: (i) Gaze direction and (ii) Head direction. Each attentional component is modeled by a set of rays in a cone that may intersect objects in the environment. We define a quantitative object attention based on the distance between the rays that intersect the 3-Dimensional bounding box and the center of the object. We implemented a pipeline that estimates VFOA with the following components: (i) Object detection in images, (ii) stereo-based estimation of 3D bounding boxes, (iii) 3D Head orientation and (iv) 3D gaze orientation.

The remainder of the document is structured as follows: Section 2 presents a brief background about previous work on object detection, gaze and head pose estimation, and finally, attention. Section 3 introduces the framework used to estimate the VFOA from the gaze and head pose estimation. Section 4 ultimately discusses the synthetic experiments and presents the results of the MPI Sintel dataset on the achieved VFOA. In Section 5, we discuss and analyze the main contributions produced and conclude the work of our findings, followed by future work.

II. Background and Related Work

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 and after VFOA was estimated it can be utilized as a primary source for human-robot interaction. For example in a multi-party dialog, gazed combined with VFOA can be used to establish a face to face communication or to indirectly inform the other person as a speech-turn taking,

This work was supported by FCT with the LARSyS - FCT Project UIDB/50009/2020, and also by the project IntelligentCare – Intelligent Multimorbidity Management System (Reference LISBOA-01-0247-FEDER-045948), and by the FCT project HAVATAR-PTDC/EEI-ROB/1155/2020

in a complement of speech communication, or simply draw someone attention.

VFOA may be simplistically defined by an imaginary line that initiates from a perceiver is face orientation, the person that we want to estimate the VFOA, to the target perceived or by the perceiver gaze direction to the target. For that reason, one may state that the VFOA of a person i is the target j when the two gaze directions are aligned, but this definition is not always true, cases where the head pose differs from the gaze. With that in mind, the VFOA can be calculated using the head orientation [4], or both the head and eyes together [5] to discover the proper gaze direction and the identification of the target.

Some of the authors do not consider VFOA estimations methods based on eye detection and eye tracking a proper and more accurate solution, as we could have imagined because until now, no gaze estimation method could deal with unconstrained scenarios, particularly the occlusion partial or total of the eyes and the inaccuracy obtain from gaze in a long distances.

In consequence, some review methods focus exclusively on head pose. One of them [6] proposed estimating the VFOA through a gaze cone fit around the head providing the orientation and projection space where we could identify the targets lying inside this cone, however, this method had several major drawbacks due to its vagueness in detecting the object inside the projection space.

Another interesting method proposes to estimate the VFOA by analysing the interaction and behavior of the participants engaged in meetings [4], which are characterized by interactions between individuals where the main forms of interaction are movements of the head and/or eyes, and speech. In this approach, a Hidden Markov Model (HMM) was proposed to infer VFOA from head and body orientations. This model has the advantage of modelling contextual information, e.g. participants tend to be more observant of the speaker, or even to the robot or the object that was referred. The results shows that the addition of contextual information improves the overall performance of the VFOA compare with methods without this contextual information, nevertheless, this turn out to be a disadvantage because we are adding new information, such as voice recognition or the speaker identification, also for HMM when we change the context of the problem the parameters of the HMM model needs to be re-estimated, e.g., it must be trained for each of the meetings lay-out.

One last method [5] was analyzed that exploits the correlation between eye gaze and head movements. Where a Bayesian switching dynamic model was used for the estimation, tracking of the gaze and VFOAs of all persons involved in the social interaction. It is assumed that the target head poses, both orientation, and location are directly achieved from the data, on the contrary, the gaze and VFOA are unknowns variables.

We note that gaze inference from head orientation is an ill-posed problem. Indeed, the correlation between gaze and head movements is person dependent as well as context dependent. It is however important to detect gaze whenever the eyes cannot be reliably extracted from images and properly analyzed. A particular feature used in this method [5] resides in the fact that eye detection is not necessary, although there are many gaze estimation methods. This feature brings advantages and disadvantages to the calculation of the VFOA itself, one advantage is the fact that most of the gaze estimation architectures can not deal with some of the unconstrained situations, so it turns the method efficient and useful in a rather wide number of situations, e.g. social interaction.

However, estimating the gaze through head orientation alone is a propagation error, because the correlation between gaze and head movements is not as linear as it was proposed in [5], it depends on the person and the situation as well.

III. 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.

III-A 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.

III-B 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. [7] describe, the field of view can be divided into several parts: the field of focus (30º), the field of vision (60º), the field of peripheral vision (120º), and the field blind to the eye.Since the angle of the field of view depends on the application, this is a parameter that needs to be tuned. We expect to fall withing the field of focus, which represents the person's amplitude of focus and solves the uncertainty in the gaze position. The scaled cone is constructed with its vertex at the middle of the eyes of the

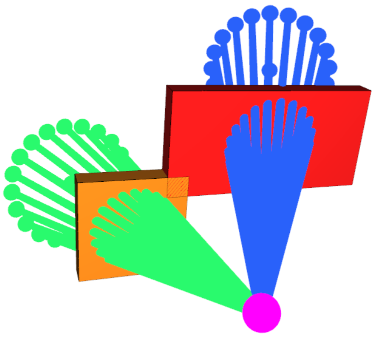


Fig. 1. Illustration of the two components of the field of view. The head pose direction points towards the red bounding box, and the gaze direction to the orange bounding box. The rays of each component are uniformly distributed on the cone surface and there is an additional ray in the symmetry axis of each cone.

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 placing 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 as shown in Figure 1).

III-C 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.

III-D 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.

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

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

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.

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

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. [8] 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.

$$\begin{cases} \omega_G = 60\% \\ \omega_{HP} = 40\% \end{cases}, \quad \text{G}\lhd\text{HP} \leq 30^\circ \tag{4}$$

$$\begin{cases} \omega_G = 90\% \\ \omega_{HP} = 10\% \end{cases}, \quad \text{G}\lhd\text{HP} > 30^\circ \tag{5}$$

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.

Variable	**Description**
P_I	Point of intersection
C_{BB}	Centroid of the Bounding Box
A_{IG}	Attention of each ray for the gaze cone
A_{IHP}	Attention of each ray for the head pose cone
A_T	Total Attention for a specific object
n_r	Number of rays in the cone
d_{min}	Minimun distance from the box to the Centroid
d_{max}	Maximun distance from the box to the Centroid
ω_G	Weight for the gaze cone
ω_{HP}	Weight for the head pose cone
G$\lhd$HP	Angle between the gaze and the head pose original vector

TABLE I
LIST OF THE VARIABLES FOR ESTIMATING THE VFOA.

IV. VFOA Architecture for stereo-based systems

In the case of humanoid-like robots that have heads and a pair of cameras that act as eyes, stereoscopic perception of the environment provides 3D location of the objects of interest as well as the persons. We aim to implement the VFOA for humanoid social robots that have stereo systems in their heads. Figure 2 shows the architecture, where the stereo reconstruction of the world provides the input for: (i) Object recognition and 3D bounding box estimation, and (ii) Head pose and gaze tracking. These two components are the needed input for VFOA estimation, which identifies the visual target, e.g., person or object, on which the person of interest is focused.

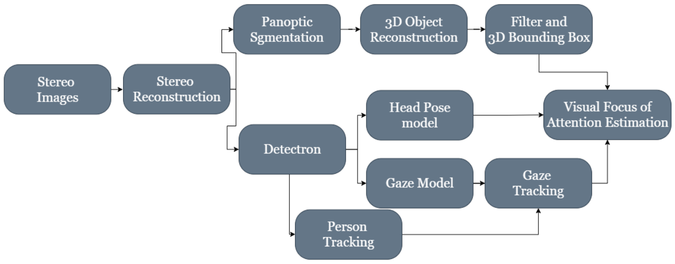


Fig. 2. Components of the VFOA architecture. Arrows indicate the information flow direction.

IV-A Stereo Reconstruction

We rely on traditional (dense) stereo algorithms such as Semi-Global-Block-Matching (SGBM) [9] and Block-Matching (BM) [10]. These algorithms require: (i) Intrinsic calibration of each camera that removes distortions and map 3D points to image pixels, (ii) extrinsic calibration of the cameras (Pose of one camera with respect to the other one). The calibrated images are the input to the disparity map estimation by triangulation [10], which computes the depth value at each pixel by minimizing the squared loss of all pixels.

IV-B Image Segmentation

Image segmentation can be divided into two mains parts: semantic segmentation, which associates each pixel of an image with a class label, and instance segmentation that masks each instance of an object in an image. A recent approach in image segmentation is Panoptic segmentation [11], which is the combines semantic and instance segmentation. Panoptic segmentation provides the location in the image of the object and its class, potentially identifying and locating the object of focus for effective interaction in one model. The output of Kirillov et al. [11] work provides two channels: one for pixel's label, which represents the semantic segmentation, and another for predicting each pixel instance.

IV-C 3D Object Reconstruction

Each object instance detected by the Pantoptic segmentation is represented by a set of connected pixels. Each set of connected pixels is reconstructed into a point cloud, using the depth map and their corresponding pose information with respect to the stereo pair. Lastly, with the instance point cloud, we can filter and estimate the 3D bounding boxes for each object/person detected.

IV-D Gaze Estimation

Gaze direction estimation is only accurate when the persons are wearing glasses that have cameras that point to their eyes. Otherwise, gaze estimation suffers from low accuracy in unconstrained scenarios. Recent methods develop deep learning architectures that rely on Convolutional Neural Networks to compute features and Long-Short Time Memory to filter estimations across time [12] have shown robustness across time and stable results. Gaze360 [12] requires as input the cropped region of the face, which is provided by the detectron2 [13] algorithm. To add robustness to the Gaze detection, we implemented tracking on both the detected faces and the Gaze direction estimation.

IV-E Person Tracking

The faces detected in the previous image are associated with the faces in the current image, using the Intersection Over Union (IOU) feature. 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.

IV-F Gaze Tracking

We implemented a Kalman Filter tracker based on the work in [14], which estimates the location and velocity of gaze assuming that are independent variables. In Toivanen's [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 not the common case. Thus, we remove the independence, computing gaze velocity as the derivative of the position.

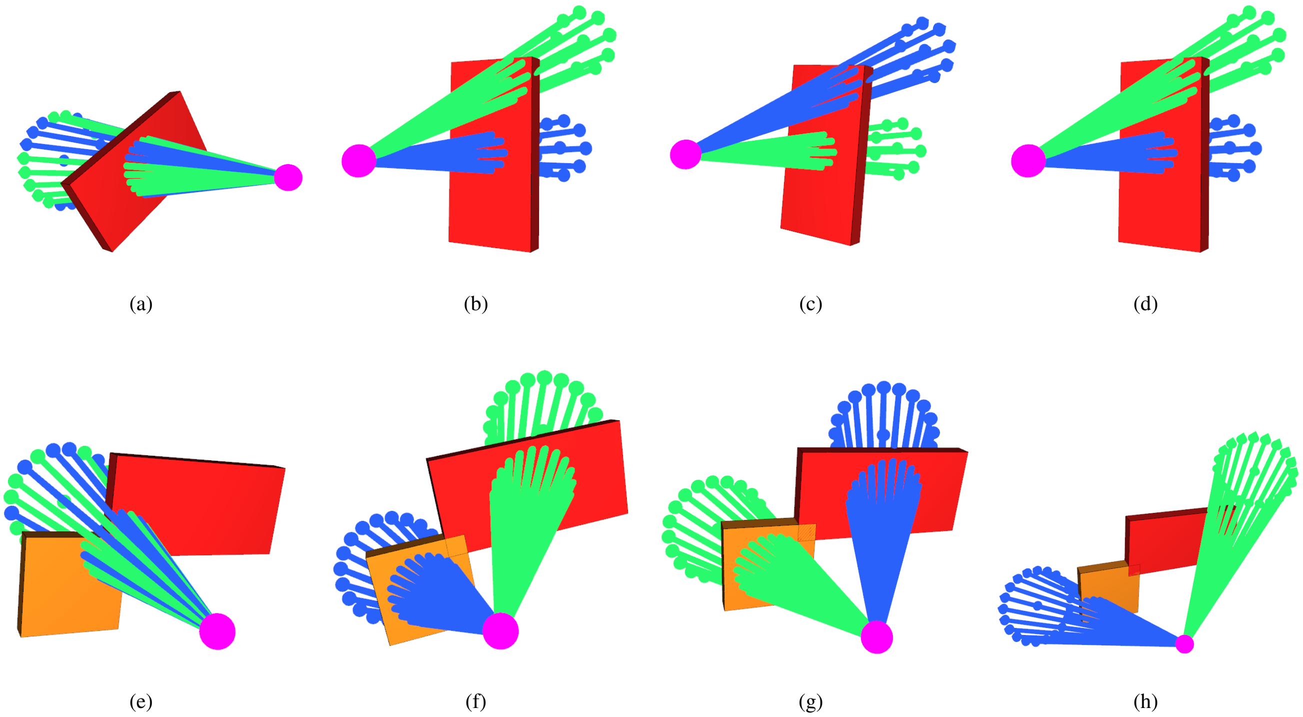
IV-G Head Pose Estimation

The model chosen for the estimation of the Head Pose task was the FSANet [15] 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.

V. Model Validation and Experimental Results

V-A VFOA parameters selection

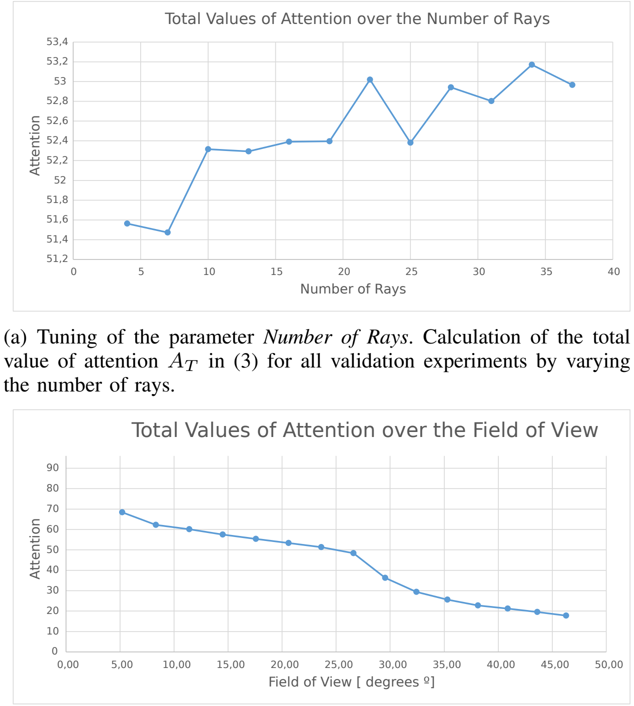
The VFOA attention values of (1)-(2) depend on the size of the Field of view and the number of rays. To select these two parameters, we run synthetic scenarios while varying the parameters. In the experiments, the Number of Rays was initialized with a value of 4 and increased until 40. The Field of View was initialized with a value of $5.2°$ (corresponding to a cone radius of $5cm$) and increased until $46.3°$ ($47cm$).



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

(e) (f) (g) (h)

Fig. 3. Scenarios with one object (top row) and two objects (bottom row). The objects are represented by their bounding boxes, the head direction by the blue cone and the gaze direction by the green cone.



(a) Tuning of the parameter *Number of Rays*. Calculation of the total value of attention A_T in (3) for all validation experiments by varying the number of rays.

(b) Tuning of the parameter *Field of View*. Calculation of the total value of attention A_T in (3) for all validation experiments by varying the cone's radius.

Fig. 4. Experiments for field of view and number of rays parameter selection

Figure 3 shows the synthetic scenarios considered to select the parameters of the head direction cone and gaze direction cone. Figure 4 shows 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. [7], 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.

V-B VFOA Experimental Results

In this section we present the estimation of the target object using the attention A_T in (3), using the parameters selected in the synthetic experiments.

V-B1 Dataset

The 3D-animated short film from the **MPI Sintel [16]** dataset, on the clip "Bandage" of stereo images is used to estimate VFOA. The interaction happens betweeen 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 5 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 was measured through the Frame-based Recognition Rate (FRR), which corresponds to the percentage of frames where the VFOA matches the ground truth label.

V-B2 Experimental Overview

We evaluate the VFOA estimation with two cone sampling options: Single-cone and multi-cone computation. In addition, 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. 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.

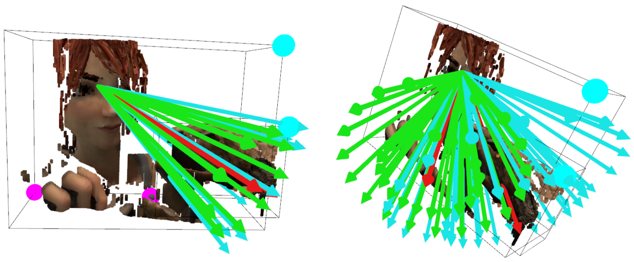


Fig. 5. On the left side, the reconstructed in 3D view with the person and the bird correctly identified by the bounding boxes, and the two cones each, for the gaze, green, and the head pose, blue. The Rays of the cones were used to calculate the VFOA. On the right side, the VFOA estimation using two cones.

V-B3 MPI Sintel Experiment With Single Cone per component

Stereo BM algorithm. Table II 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%. In this case, 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.

Stereo SGBM algorithm. The target object 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. Similar to the BM stereo algorithm, the head pose continues to have a more prominent influence than the gaze on the final attention results.

V-B4 MPI Sintel Experiment with two cones per component

In the previous experiment, the attention algorithm only used one cone to calculate the value of attention. We add one more cone while having small and medium radii. Both radius values showed the best results and matched the definition of the focus field of view. Right side of Figure 5 displays the two cones and the characters' interaction in the MPI Sintel ”Bandage” dataset.

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 BM case), by having doubled the number of rays.

Stereo SGBM algorithm. The target object 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. 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. In Figure 6 we 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 II
SUMMARY OF THE VFOA EXPERIMENTAL RESULTS.

VI. CONCLUSIONS

This work presents an approach to estimate the VFOA based on the gaze and head orientation. Specifically, we introduced

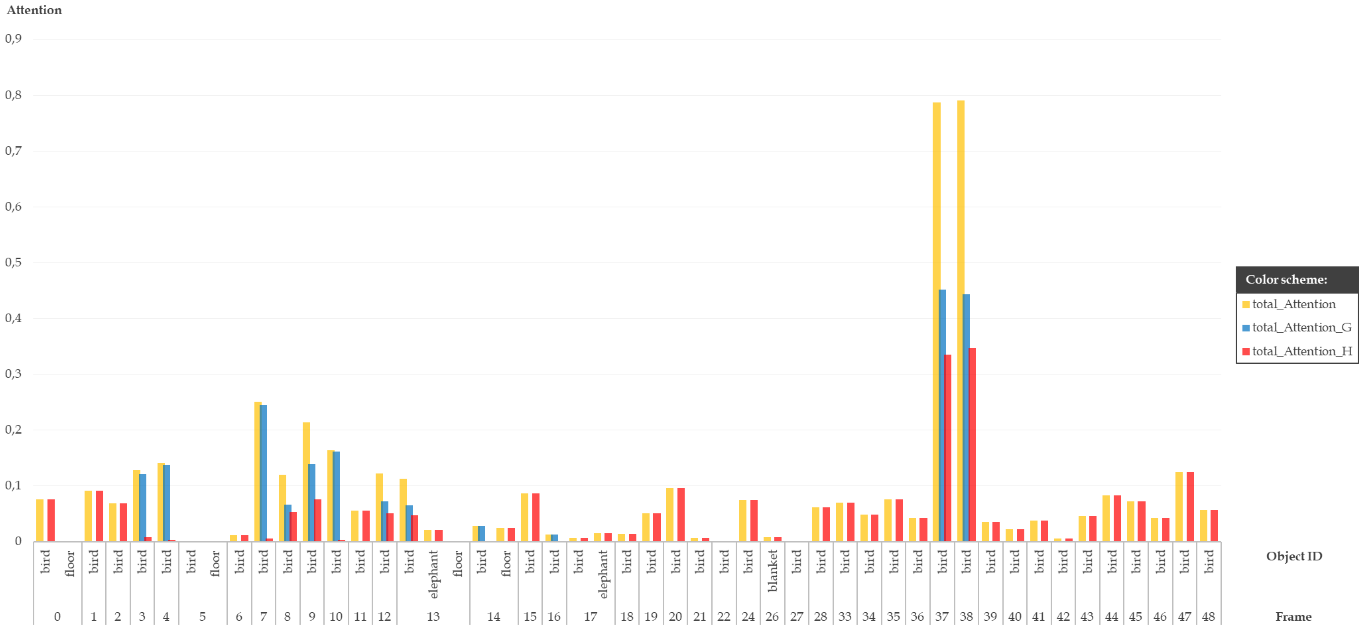


Fig. 6. MPI sintel “Bandage” video reconstructed in 3D view with the person and the bird correctly identified by the bounding boxes, and the two cones each, for the gaze, green, and the head pose, blue. The Rays of the cones were used to calculate the VFOA.

a target attention model that weighs gaze direction and head orientation. Each component is modeled by a set of rays that lie in the surface of a cone. The field of view (cone angle) and number of rays are the main parameters, which were selected on a set of synthetic experiments. The objects in the scene are modeled by a 3-dimensional bounding box that is obtained from panoptic segmentation and stereo reconstruction.

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. Our 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. The main drawback of our approach is not providing the results in real-time.

References

[1] C. Breazeal, “Emotion and sociable humanoid robots,” *International Journal of Human-Computer Studies*, vol. 59, pp. 119–155, 07 2003.

[2] B. Ogden and K. Dautenhahn, “Robotic etiquette: Structured interaction in humans and robots,” *Proc. SIRS2000, Symposium on Intelligent Robotic Systems, Reading, UK*, no. 1998, pp. 353–361, 2000.

[3] 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)*, vol. 1614, pp. 761–768, 1999.

[4] S. O. Ba and J. M. Odobez, “Recognizing visual focus of attention from head pose in natural meetings,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 1, pp. 16–33, 2009.

[5] B. Masse, S. Ba, and R. Horaud, “Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2711–2724, 2018.

[6] M. J. Marín-Jiménez, A. Zisserman, M. Eichner, and V. Ferrari, “Detecting people looking at each other in videos,” *International Journal of Computer Vision*, vol. 106, pp. 282–296, 2013.

[7] 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.

[8] 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*, vol. 45, no. 5, pp. 1019–1028, 2007.

[9] H. Hirschmuller, “Stereo processing by semiglobal matching and mutual information,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[10] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library.* ” O’Reilly Media, Inc.”, 2008.

[11] A. Kirillov, R. B. Girshick, K. He, and P. Dollár, “Panoptic feature pyramid networks,” *CoRR*, vol. abs/1901.02446, 2019. [Online]. Available: http://arxiv.org/abs/1901.02446

[12] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, “Gaze360: Physically Unconstrained Gaze Estimation in the Wild,” 2019. [Online]. Available: http://arxiv.org/abs/1910.10088

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

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

[15] 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.

[16] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, “A naturalistic open source movie for optical flow evaluation,” in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 611–625.