SCENE-pathy: Capturing the Visual Selective Attention of People Towards Scene Elements

SCENE-pathy: Capturing the Visual Selective Attention of People Towards Scene Elements

Andrea Toaiari[1], Federico Cunico[1], Francesco Taioli[1], Ariel Caputo[1], Gloria Menegaz[1], Andrea Giachetti[1], Giovanni Maria Farinella[2], and Marco Cristani[1]

[1] University of Verona, Verona, Italy
name.surname@univr.it
[2] University of Catania, Catania, Italy
gfarinella@dmi.unict.it

Abstract. We present *SCENE-pathy*, a dataset and a set of baselines to study the visual selective attention (VSA) of people towards the 3D scene in which they are located. In practice, VSA allows to discover which parts of the scene are most attractive for an individual. Capturing VSA is of primary importance in the fields of marketing, retail management, surveillance, and many others. So far, VSA analysis focused on very simple scenarios: a mall shelf or a tiny room, usually with a *single* subject involved. Our dataset, instead, considers a *multi-person* and much more complex 3D scenario, specifically a high-tech fair showroom presenting machines of an Industry 4.0 production line, where 25 subjects have been captured for 2 minutes each when moving, observing the scene, and having *social interactions*. Also, the subjects filled out a questionnaire indicating which part of the scene was most interesting for them. Data acquisition was performed using Hololens 2 devices, which allowed us to get ground-truth data related to people's tracklets and gaze trajectories. Our proposed baselines capture VSA from the mere RGB video data and a 3D scene model, providing interpretable 3D heatmaps. In total, there are more than 100K RGB frames with, for each person, the annotated 3D head positions and the 3D gaze vectors. The dataset is available here: https://intelligolabs.github.io/scene-pathy.

Keywords: Visual Attention · Social Signal Processing · Gaze Estimation · Benchmark

1 Introduction

Capturing the attention of people toward specific elements in a scene is an attractive yet unsolved problem in computer vision. It is a precious skill in practical fields such as marketing and retail management [5]: knowing where the attention of most people will be directed allows one to properly set up the advertisements inside a building, as well as to charge adequately for the posting of the advertisements themselves. This type of attention is called *visual selective attention* (VSA), *i.e.*, the process of directing the gaze to relevant visual stimuli while ignoring the irrelevant ones in the environment [8]. VSA is generally studied

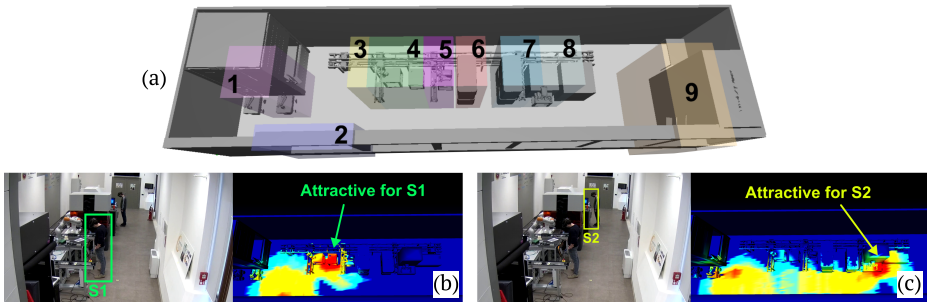

Fig. 1. (a) The 3D model of the *SCENE-pathy* showroom scenario, with the 9 possible areas of interest highlighted by colored bounding boxes. (b, c) On the left is an input video frame, in which we highlight a single subject. On the right, the corresponding VSA estimation is in the form of a 3D map, where hotter colors indicate areas that attracted more VSA.

by exploiting the eye gaze dynamics [7,21] in very simple 2D or 3D scenarios, where the area of interest is limited to planar scenes, such as mall shelf [4,32], or inside tiny rooms [29,16]. This is a real limitation, as visual selective attention is certainly important even in more structured and larger scenes. In this paper, we present *SCENE-pathy*, a dataset collected in a 20 by 5 meters showroom of a tech fair (Fig. 1), where 9 different work areas are distinguishable, each one equipped with costly hardware to be exhibited. The showroom has been manually reconstructed beforehand as a 50K-points 3D cloud. A total of 25 unacquainted subjects have agreed to participate to build the dataset. For each participant, we have collected video sequences and sensory data, thanks to the use of Hololens 2 devices, for about 1.5 hours of data footage. Hololens 2 allowed us to extract tracklets of 3D head positions and eye gaze trajectories, which we collected as annotations for the dataset.

As an additional, higher-level, annotation, the subjects filled out a simple questionnaire, communicating which parts of the scenes attracted their attention the most. A dataset with these characteristics would be a novel and useful asset to study problems related to VSA estimation and attention target detection in general, both in the 2D and 3D spaces. To demonstrate some applications on the dataset, we propose a set of baselines to capture the VSA of single and pairs of people. In general, these baselines individuate and track individuals by 3D pose estimation, using solely the RGB video stream as input. Successively, they intersect a physiologically plausible view frustum, fitted on the estimated head pose, with the 3D point cloud (+ 3D culling) of the scene. The resulting intersections are considered as a vote of VSA. Across time, votes are accumulated, providing a 3D weighted map that, for each person, indicates which parts of the scene have attracted their visual selective attention the most (see Fig. 1b, Fig. 1c). The use of a coloured 3D point cloud heatmap is a novel concept in this type of research. The baselines are enriched by social signal processing (SSP) findings [4,31], merging computer vision techniques with social sciences. It is

widely known that while moving the visual attention is aligned preferably with the motion vector [15]. Furthermore, social interactions require visual attention, since humans are naturally interested in observing the other interactants to capture their body language [6]. Put simply, to deal with these aspects in a principled way, our baselines weigh less VSA votes when the person is moving or engaging in social interactions.

In summary, our contributions are the following: *(i)* we introduce *SCENE-pathy*, a dataset with more than 100K RGB frames with associated, for each person, the 3D head positions and the 3D gaze vectors captured by Hololens 2 devices; *(ii)* a set of baselines capturing VSA from RGB frames in a complex multi-person scenario, taking into account also the social signal processing point of view; *(iii)* a new way of encoding the visual attention using a coloured 3D point cloud, manually reconstructed to allow the VSA counting even on areas occluded to the camera.

2 Related Works

Attention target detection. The idea of continuously estimating attention on a complex 3D heatmap from a third-person perspective is a novel task in the literature. The most similar task is attention target detection [9,32,2,10], or gaze-following [27,16,22], which aims at identifying the object looked at by a certain person in an image or video and it is useful to identify intentions and predict future actions. In [27,32,9], the task is tackled in the 2D image space, exploiting insights from the saliency maps theory [17]. The dataset proposed in [27], despite being one of the most used, presents shortcomings when considered from a 3D standpoint, such as the inability to counteract the effect of occlusions and perspective. Some recent works extend their models to handle 3D depth estimation [2,16,10,22], trying to figure out the distance from the camera of the various objects in the scene. In [2], 3D point clouds are constructed directly from 2D images and a model incorporates both the scene contextual cues and the predicted probabilities to refine the target fixation prediction. In [16], two datasets for the 3D gaze-following task are proposed, but the subject's sightlines were annotated manually, which renders them potentially inaccurate. A special case is addressed in [22], where a method is proposed to find the target of attention in 360-degree images. In [29], a single-person scenario in a lab setting is presented. A depth camera extracts the 3D human skeleton to estimate the head view vector and infer the attended object of interest. However, at least one RGB-D camera is necessary and multi-person capabilities are not tested. Our proposed dataset differs from other works by leveraging a complete 3D model of the environment, enabling us to compute attention even on parts of the scene occluded to the camera. We provide both low-level annotations, which are 3D positions and gaze vectors obtained from top-tier devices, and high-level annotations in the form of interest questionnaires, filled by the subjects after visiting the scene. Additionally, we capture VSA in a multi-person setting, through a top-down pose estimation pipeline, starting from the mere RGB video stream.

Visual attention mapping. One of the distinctive features of the proposed method is the use of the environment point cloud as a heatmap, with a coloring scheme that shows the areas of greatest interest. In [3], the analysis of the visual attention in a museum is carried out with a simple subject-artwork association, without considering the environment as a whole. The works of [28] and [4] exploit an external point of view of the scene and propose the concept of a colored attention heatmap applied to the environment. Differently from us, the used maps are not fine-grained, the frameworks did not consider the interplay of motor activities nor social interactions, and discriminate only a fixed discrete number of head orientations, while we have continuous orientations.

The Social Signal Processing Point of View. Humans sample the visual world by making eye movements to direct a centralized region of visual acuity towards different parts of the environment [23]. This allows to direct awareness to relevant stimuli while ignoring irrelevant ones in the environment and is called visual selective attention [14]. Eye gazing is the clearest way to capture the VSA [12]. In the absence of eye-tracking capabilities due to low-resolution images, the head pose is the second best indicator of where eyes may be fixated [24]. This happens due to the orbital reserve [13], the tendency for the eyes to be positioned centrally in the orbits, and the fact that people orient their heads towards important features in the environment [12]. Selective attention and vigilance have important interplay during activities such as walking [11] and interacting with other people [31]: it is known that without walking or the presence of social interaction, eye gazing is most probably associated with visual selective attention [14]. We, therefore, propose to incorporate these findings in our approach, demonstrating that they can help reduce the number of false positives when estimating the VSA.

3 The *SCENE-pathy* Dataset

SCENE-pathy is the first dataset for VSA estimation that features accurate 3D annotations of the position and gaze direction of multiple people moving in the scene, acquired automatically by head-mounted Hololens 2 devices. The main goal is to provide a novel benchmark to study whether VSA can be captured in a non-collaborative scenario using computer vision techniques on a monocular video stream, taking into account also the SSP point of view.

To this end, we captured two kinds of ground truth data: *low-level* and *high-level*. We define low-level data as the annotations for the 3D head position and gaze vector of the subjects, relative to a fixed starting position, captured using Hololens 2 devices (one for each person). These ground truth data were then synchronized with the RGB video frames, captured from a single wall-mounted camera, and the starting position is shifted according to the environment coordinate reference system. Since the main challenge we want to explore in this paper lies in capturing perceptual-cognitive processes, *i.e.*, how SSP findings can help reduce false positives in the VSA estimation, we also introduce high-level

ground truth in the form of simple questionnaires, in which the subjects have to rank, in descending order of perceived interest, all the 9 stations in the scene. The scene is a $20l \times 5d \times 4h$ meters showroom with 9 different industrial stations, replicating a modern Industry 4.0 production line (see Fig. 1a).

We manually created an accurate 3D model of the environment using Blender, sampled it into an equally distributed point cloud, and defined 3D bounding boxes around each machine. This will be useful to measure the final VSA of each object. The biggest station measures, in meters, $2.5l \times 3d \times 3.5h$, and the smallest measures $0.3l \times 0.1d \times 0.6h$. The dataset is organized into two experimental sessions: *unsupervised* and *supervised*, depending on whether the subjects freely explored the environment or were directed to a specific industrial station. Note that the supervised experiments were performed after the unsupervised ones. Each session was further divided into *single-person* and *multi-person*. The multi-person experiments were conducted with two people at the same time, in which we require them to engage in a discussion, hence having a social interaction.

Unsupervised experiments: In the single-person unsupervised experiments, the subjects were asked to freely explore the environment and then fill out the questionnaire. In the multi-person variant, the couple was also expected to discuss their station of interest, motivating their choices, before filling out the questionnaire. The average running time of each experiment is about 2 minutes.

Supervised experiments: In the supervised experiments, for each subject, we selected *beforehand* a specific industrial station to inspect. Obviously, the subjects are now more familiar with the environment, as they have already seen it in the unsupervised session. In this case, we want to be sure which station should be the most looked at, to see if we can estimate it correctly. As in the unsupervised session, for the multi-person variant, we also asked the subjects to have a brief discussion after the exploration. Since the subjects can't undergo the unsupervised trials more than once, this second type of test is useful to further enrich the dataset and provide simpler and less randomized movement patterns.

The *SCENE-pathy* dataset contains 19 unsupervised experiments (13 single-person and 6 with pairs), and 60 supervised experiments (48 single-person and 12 with pairs). For the data collection, we hired 25 unacquainted subjects: 23 males and 2 females, average age of 27. The dataset is composed of the 3D model of the scene and more than 100K RGB frames with associated, for each person, the accurate annotation of 3D head positions, the 3D gaze vectors, and also the questionnaires as additional annotation of the actual interest of the people.

4 The Proposed Baselines

In this section, we present a modular baseline dubbed *Socio-Dynamic VSA* (*SD-VSA*) to compute the VSA on the *SCENE-pathy* dataset. Furthermore, to demonstrate the effectiveness of the involved social signal modules, we provide simpler ablative versions. The first module of the pipeline estimates the 3D pose of each subject, in the form of a 3D skeletal model. The captured videos are processed by a top-down 2D pose estimation algorithm [30], with a tracking

module [33] to maintain the subject ids. The extracted 2D pose joints are then converted by a 3D pose lifter [25] to 17-joint 3D skeletal models, centered initially at the axis origin. These skeletons are then located in their correct scene locations using a homography matrix to transform points from the camera space to a top-view planimetry of the environment. The vector passing through the two 3D joints on the head, one for the nose and the other between the ears, becomes the central axis of a pyramid-shaped view frustum. This represents the direction and area of focus of the subject's gaze. As shown in Sec. 5, this estimated vector aligns sufficiently well with the ground truth sensor provided by the Hololens 2 devices, in terms of both pan and tilt angular errors.

At the start, each of the N 3D points of the point cloud representing the environment, dubbed P, is assigned a null attention value. For each frame t, and for each human subject i, the head position and the derived gaze direction are used to estimate the subset of visible points F_t^i. This is done in two steps: first, the occluded points are determined and removed using the Hidden Point Removal operator [18]. Then, a view frustum culling is performed considering a horizontal FOV of 90° and a vertical FOV of 60°, mainly based on [20]. In the last step, the attention values of the non-occluded points inside the frustum are increased according to a weighting scheme, with a maximum around the central axis and exponentially decreasing at the margins, following the idea of the attention spotlight model [26]. The view frustum is quantized into three concentric parts with different weights to speed up the computation.

Specifically, the VSA_t^i for a specific frame t and a subject i is composed by the indexes of the point cloud with an associated attention value $s_{t,k}^i = \gamma$, if $p_k \in F_t^i$, else 0, with p_k the k-th point of P. The weight γ is quadratic and it is chosen depending on which subdivision of the view frustum contains the considered point. The weights can then be further modulated according to the movement and the social activity of the subject, as described in the following subsections. Considering the interval of frames T and the subject i, the individual map VSA^i is the sum of the map scores at each frame t: $\text{VSA}^i = \sum_t^T \text{VSA}_t^i$.

The Dynamic Module. Sec. 2 explains that motor activity and social interaction are overriding the visual attention a subject may spend towards the environment [11,12]. In the *Dynamic* module of *SD-VSA*, the motor activity is modeled by changing the equation for $s_{t,k}^i$, substituting γ with $(\gamma \cdot w(v))^2$, where $w(v)$ is a weight inversely proportional to the velocity of the subject in the scene. In simple words, the higher the velocity, the lower the weight that a certain intersection between the view frustum and the scene does have.

The Social Interaction Module. This module copes with the fact that social interactions inhibit the visual selective attention of the individuals towards the scene elements [6,31]. Many approaches are available in the literature; the most effective ones take into account proxemics cues (the position of the interactants) and kinesics cues (body poses). In particular, [4] individuates a social interaction when people are close enough (accounting to Kendon's findings on Hall's usage of

the personal space [19]). In practice, people at a distance of less than 1.2 meters, that are pointing their faces at each other, individuate a social interaction.

All these cues are available: thanks to the multi-person pose estimation, this configuration can be found simply by analyzing each possible pair of stationary individuals. On each pair, if the people are close enough, two conditions are checked: if there is an intersection between the view frustums and if each one of the vision cones contains the other subject's head. When in the presence of both conditions, the counting of the scene-frustum intersections is suspended.

Ablative Baselines. Alternative baselines can be obtained by suppressing either one or both of the social signal modules. The *Vanilla* baseline indicates the scenario in which none of the modules is active, and the VSA maps are always increased by the same amount, no matter the speed of the subjects or the occurrence of social interactions. *DYN* and *SOCIAL* represent respectively the baseline in which only the *Dynamic* module is active or only the *Social* module is active. Of course, the *Social* module only makes sense in experiments where more than one subject is present.

5 Experiments and Results

In this section, we present the experiments and comment on the results obtained by applying our baselines. Firstly, we show that our 3D head pose estimation pipeline (approximating the gaze) performs sufficiently well, demonstrating quantitatively the orbital reserve effect [13] discussed in Sec. 2. Since all of the baselines of Sec. 4 depend on the gaze direction, this is a necessary step of the analysis. Then, we show the capability of *SD-VSA* and the alternative baselines in providing accurate VSA maps on the *SCENE-pathy* dataset, despite the uncertainty in the true gaze direction. Finally, we show that our baseline can provide useful VSA maps also in more crowded scenarios, using the GVEII benchmark [1].

Capturing the Human Gaze. We extracted the Hololens 2 head poses and compared them with the eye-tracking data by measuring the pan and tilt angles error, resulting in less than 1° and 3° discrepancy, respectively. Then, to assess the quality of our head pose estimation pipeline we calculated the pan and tilt error of the whole dataset in the range [0°, 360°] for the pan and [0°, 180°] for tilt. For pan, 70% of the elements are in the range of 0° and 180°, while on tilt 97% of the elements are between 60° and 135°. Globally, the average pan error is 52.36° and the average tilt error is 50.81°. We consider these results acceptable, given the size and complexity of the analyzed scenario.

Unsupervised Experiments. In this section, we show how VSA maps associated with each subject are coherent w.r.t. the high-level ground truth questionnaires discussed in Sec. 3, and the impact of the *Dynamic* and *Social* modules detailed in Sec. 4. We start by comparing the ranking from the questionnaires with the ones obtained by the scores of the computed VSA. In particular, the

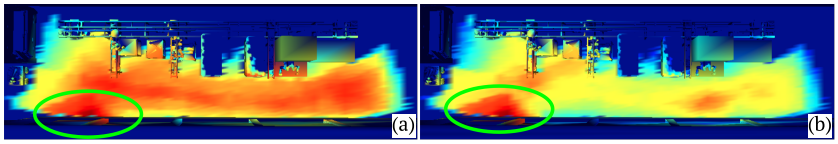

Fig. 2. Qualitative results of an unsupervised single-subject experiment. (a) *DYN* module disabled (b) *DYN* module active. Without the *DYN* module, the resulting VSA is definitely noisier. The target object was an advertising board, circled in green.

Table 1. The accuracy in unsupervised experiments.

Method	**CMC Single-Person**				**CMC Multi-Person**			
	Rank-1	Rank-2	Rank-3	Rank-4	Rank-1	Rank-2	Rank-3	Rank-4
Vanilla	30.8	46.2	61.5	76.9	25.0	41.7	50.0	66.7
DYN	38.5	61.5	84.6	100.0	25.0	50.0	75.0	83.3
SOCIAL	-	-	-	-	33.3	58.3	75.0	100.0
SD-VSA	-	-	-	-	33.3	58.3	83.3	100.0

9 objects are ranked taking the maximum score within the bounding boxes defined around each station. Subsequently, we compute over all the experiments the Cumulative Matching Curve (CMC), comparing the two rankings. This curve defines the probability of a specific object having the highest score, and hence being the correct object, within the first N points of the curve. We first analyze the *unsupervised single-person* sequences. Considering the presence of a single person, we evaluated the system in two configurations: one with the *DYN* module and one without (*Vanilla*). Numerically, the *DYN* module gives an accuracy boost, and Fig. 2 shows that VSA maps generated without it are definitely noisier, showing a marked trace on the floor where the subject has passed.

In the *unsupervised multi-person* sequences, together with the *Vanilla* and *DYN* module we also consider whether the accumulation on the maps is stopped by the occurrence of social interactions, dubbed *SOCIAL*. This leads to four possible combinations (see Tab. 1). As a metric, we report the rank 1-4 accuracies, which show the predominance of the complete *SD-VSA* model as expected.

Discussion: The single-person unsupervised experiments results are presented in Tab. 1. Since the subjects have never seen the environment, they usually start by navigating the whole hallway, before focusing on some specific industrial station. The accuracy of the system solely relying on people's location is low, achieving below 50% at rank-2, which is improved to over 60% introducing the *DYN* module. The significant difference in accuracy between rank-1 and rank-2 can be explained by the fact that the machines are distributed very close to each other, indicating that often the correct object is spatially close to the estimated one. Moreover, since our pipeline is modular, replacing the gaze estimation module with a more accurate one could certainly provide an additional boost in the performance described by the CMC curve.

Multi-person unsupervised experiments also present a non-trivial task to solve, as the presence of two individuals can introduce noise in the VSA estima-

Table 2. The accuracy in supervised experiments. In this case, the value tells us if we were able to identify the specific station assigned to subjects.

Method	Single-Person	Multi-Person
Vanilla	56.3	50.0
DYN	62.5	54.2
SOCIAL	-	62.5
SD-VSA	-	70.8

tion, due to social interactions. Incorporating an SSP module helps to improve the performance of the system. It is important to note that the results of multi-person experiments are generally lower than those of single-person ones, since the pose estimation pipeline face challenges with occlusion and identity swapping caused by multiple individuals exploring the large scene. Finally, the actual interest of the subjects, collected through the questionnaires, shows that solely relying on gaze estimation is not enough, and insights from SSP increase the chance of capturing the actual interest.

Supervised Experiments. In the supervised experiments the goal is to check whether the VSA maps can highlight the object in the scene that the subject was assigned to. In this case, the ground truth is therefore the assigned station. As an accuracy measure, we consider a success whether the peak of the VSA map is inside the bounding box surrounding the target object, 0 otherwise. The results of these experiments are presented in Tab. 2. As a qualitative result, Fig. 3 shows the VSA maps after a long social interaction has occurred in the hallway. Without the *SOCIAL* module, there are incorrect attention peaks: on the wall behind subject 1 (Fig. 3c) and on the machine behind subject 2 (Fig. 3a). Instead, with *SD-VSA*, we can recognize this interaction and suspend the attention weighting, focusing only on the real attention of the subjects.

Discussion: In these experiments, the subjects are now more familiar with the environment. Additionally, since we assigned beforehand the target industrial machine, subjects naturally perform less to no exploration of the room, heading straight for the given machine. This explains why the supervised experiments achieve higher results than the unsupervised ones. The results shown in Tab. 2 clearly reflect this difference between the two sets of experiments, providing further evidence of the usefulness of taking into account SSP insights.

5.1 Does the *SD-VSA* Scale Up to Crowded Scenarios?

To prove that the SSP modules of the *SD-VSA* baseline are effective and generalize to highly crowded scenes, we use the GVEII benchmark [1], where hundreds of people walk in an outdoor mall. Positions and group formations are already available as annotations. We compute the VSA with *SD-VSA* and the *Vanilla* baseline, using a simple 3D model and projecting the frustums on the floor for visualization purposes. Note that here the view frustums are not computed through 3D pose estimation; instead, we use the oriented velocity vector derived

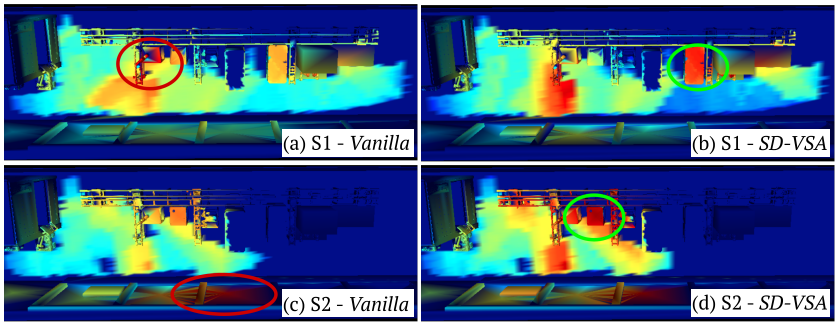

Fig. 3. Qualitative results of a supervised experiment with a pair of subjects (S1 and S2), that had a dialogue in the center of the hallway. In (b) and (d) the objects circled in green correspond to the correct most attractive points for those VSA maps. (a) and (c) show instead the false positives targets (circled in red) obtained by disabling the *DYN* and *SOCIAL* modules.

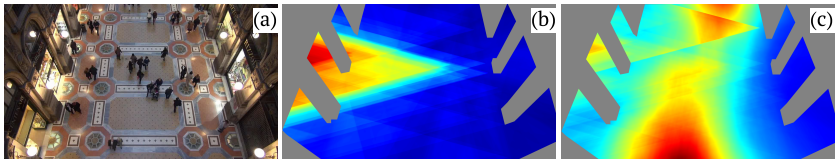

Fig. 4. (a): a frame from the GVEII dataset; (b): *SD-VSA* baseline; (c): *Vanilla* baseline: *DYN* and *SOCIAL* modules are switched off.

from the ground truth trajectory data. In Fig. 4b, it is possible to note a clearly interpretable global VSA, communicating that the people were interested in the two shop windows on the left, while the bottom-right window shop, which is closed, gather the least attention. In Fig. 4c results are incorrect, since the most interesting part of the scene appears to be the pathway, due to the vertical flow of the people.

6 Conclusions

In this paper, we propose *SCENE-pathy*, a benchmark dataset to study the visual selective attention of people towards a complex 3D scene. Along with the RGB video streams, we provide complete annotations of the 3D position and gaze direction of the involved subjects, extracted from Hololens 2 devices, and questionnaires reporting the most interesting areas visited by them. The proposed baselines, created joining pose estimation techniques and social signal processing insights, allowed us to demonstrate that more accurate VSA maps can be obtained by considering the subjects' movement speed and the possible occurrence of social interactions in multi-person scenarios.

Acknowledgments. This work was partially supported by the Italian MIUR within PRIN 2017, Project Grant 20172BH297: I-MALL - improving the customer experience

in stores by intelligent computer vision and PNRR research activities of the consortium iNEST (Interconnected North-Est Innovation Ecosystem) funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.5 – D.D. 1058 23/06/2022, ECS_00000043). This manuscript reflects only the Authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

1. Bandini, S., Gorrini, A., Vizzari, G.: Towards an integrated approach to crowd analysis and crowd synthesis: A case study and first results. Pattern Recognition Letters **44**, 16–29 (2014)
2. Bao, J., Liu, B., Yu, J.: Escnet: Gaze target detection with the understanding of 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14126–14135 (2022)
3. Bartoli, F., Lisanti, G., Seidenari, L., Del Bimbo, A.: User interest profiling using tracking-free coarse gaze estimation. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 1839–1844. IEEE (2016)
4. Bazzani, L., Cristani, M., Tosato, D., Farenzena, M., Paggetti, G., Menegaz, G., Murino, V.: Social interactions by visual focus of attention in a three-dimensional environment. Expert Systems **30**(2), 115–127 (2013)
5. Becattini, F., Becchi, G., Ferracani, A., Bimbo, A.D., Presti, L.L., Mazzola, G., Cascia, M.L., Cunico, F., Toaiari, A., Cristani, M., et al.: I-mall an effective framework for personalized visits. improving the customer experience in stores. In: Proceedings of the 1st Workshop on Multimedia Computing towards Fashion Recommendation. pp. 11–19 (2022)
6. Birmingham, E., Bischof, W.F., Kingstone, A.: Gaze selection in complex social scenes. Visual cognition **16**(2-3), 341–355 (2008)
7. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE transactions on pattern analysis and machine intelligence **35**(1), 185–207 (2012)
8. Carrasco, M.: Visual attention: The past 25 years. Vision research **51**(13), 1484–1525 (2011)
9. Chong, E., Wang, Y., Ruiz, N., Rehg, J.M.: Detecting attended visual targets in video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5396–5406 (2020)
10. Fang, Y., Tang, J., Shen, W., Shen, W., Gu, X., Song, L., Zhai, G.: Dual attention guided gaze target detection in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11390–11399 (2021)
11. Fotios, S., Uttley, J., Cheal, C., Hara, N.: Using eye-tracking to identify pedestrians' critical visual tasks, part 1. dual task approach. Lighting research & technology **47**(2), 133–148 (2015)
12. Foulsham, T., Walker, E., Kingstone, A.: The where, what and when of gaze allocation in the lab and the natural environment. Vision research **51**(17), 1920–1931 (2011)
13. Fuller, J.H.: Eye position and target amplitude effects on human visual saccadic latencies. Experimental Brain Research **109**(3), 457–466 (1996)
14. Gordon, R.D.: Selective attention during scene perception: Evidence from negative priming. Memory & cognition **34**(7), 1484–1494 (2006)

15. Hasan, I., Setti, F., Tsesmelis, T., Del Bue, A., Galasso, F., Cristani, M.: Mx-lstm: mixing tracklets and vislets to jointly forecast trajectories and head poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6067–6076 (2018)
16. Hu, Z., Yang, D., Cheng, S., Zhou, L., Wu, S., Liu, J.: We know where they are looking at from the rgb-d camera: Gaze following in 3d. IEEE Transactions on Instrumentation and Measurement **71**, 1–14 (2022)
17. Itti, L., Koch, C.: Computational modelling of visual attention. Nature reviews neuroscience **2**(3), 194–203 (2001)
18. Katz, S., Tal, A., Basri, R.: Direct visibility of point sets. In: ACM SIGGRAPH 2007 Papers. p. 24–es. SIGGRAPH '07, Association for Computing Machinery, New York, NY, USA (2007)
19. Kendon, A.: Conducting interaction: Patterns of behavior in focused encounters, vol. 7. CUP Archive (1990)
20. Kress, B.C.: Digital optical elements and technologies (edo19): applications to ar/vr/mr. In: Digital Optical Technologies. vol. 11062, pp. 343–355 (2019)
21. Li, Y., Liu, M., Rehg, J.: In the eye of the beholder: Gaze and actions in first person video. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
22. Li, Y., Shen, W., Gao, Z., Zhu, Y., Zhai, G., Guo, G.: Looking here or there? gaze following in 360-degree images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3742–3751 (2021)
23. Melcher, D.: Visual stability. Philosophical Transactions of the Royal Society B: Biological Sciences **366**(1564), 468–475 (2011)
24. Parks, D., Borji, A., Itti, L.: Augmented saliency model using automatic 3d head pose detection and learned gaze following in natural scenes. Vision research **116**, 113–126 (2015)
25. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
26. Posner, M.I., Snyder, C.R., Davidson, B.J.: Attention and the detection of signals. Journal of experimental psychology: General **109**(2), 160 (1980)
27. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? Advances in neural information processing systems **28** (2015)
28. Reid, I., Benfold, B., Patron, A., Sommerlade, E.: Understanding interactions and guiding visual surveillance by tracking attention. In: Computer Vision–ACCV 2010 International Workshops, Queenstown, New Zealand, November 8-9, 2010, Revised Selected Papers, Part I 10. pp. 380–389. Springer (2011)
29. Shi, X., Yang, Y., Liu, Q.: I understand you: Blind 3d human attention inference from the perspective of third-person. IEEE Transactions on Image Processing **30**, 6212–6225 (2021)
30. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
31. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and vision computing **27**(12), 1743–1759 (2009)
32. Wang, B., Hu, T., Li, B., Chen, X., Zhang, Z.: Gatector: A unified framework for gaze object prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19588–19597 (2022)
33. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)