EyeTrackUAV2: a Large-Scale Binocular Eye-Tracking Dataset for UAV Videos

Anne-Flore Perrin, Vassilios Krassanakis, Lu Zhang, Vincent Ricordel, Matthieu Perreira, Olivier Le Meur

MDPI

Article

EyeTrackUAV2: a Large-Scale Binocular Eye-Tracking Dataset for UAV Videos

Anne-Flore Perrin [1], Vassilios Krassanakis [2], Lu Zhang [3], Vincent Ricordel [2], Matthieu Perreira Da Silva [2], and Olivier Le Meur [1]

[1] Univ Rennes, CNRS, IRISA, 263 Avenue Général Leclerc, 35000 Rennes, France; anne-flore.perrin@irisa.fr, olemeur@irisa.fr
[2] Polytech Nantes, Laboratoire des Sciences du Numérique de Nantes (LS2N), Université de Nantes, 44306 Nantes CEDEX 3, France; krasvas@uniwa.gr, matthieu.perreiradasilva@univ-nantes.fr, Vincent.Ricordel@univ-nantes.fr
[3] Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, 35000 Rennes, France; lu.ge@insa-rennes.fr
* Correspondence: anne-flore.perrin@irisa.fr; Tel.: +33-299-84-25-73 (A-F.P.)

Version December 3, 2019 submitted to Journal Not Specified

Abstract: Unmanned Aerial Vehicles (UAVs) achieved a lot of momentum through their fast and tremendous evolution over the last decade. A multiplication of applications results from the use of the UAV imagery in various fields such as military and civilian surveillance, delivery services, and wildlife monitoring. Combining UAV imagery with study of dynamic salience further extends the number of future applications. Indeed, considerations of visual attention open the door to new compression, retargeting, and decision-making tools. To conduct such studies, in this era of big data and deep learning, we identified the need for new large-scale eye-tracking datasets for visual salience in UAV content. To address this need, we introduce here the dataset *EyeTrackUAV2* consisting of the collection of binocular gaze information through visualization of UAV videos for both free viewing and task-based attention conditions. An analysis of collected gaze positions provides recommendations for visual salience ground-truth generation. It also sheds light upon variations of saliency biases in UAV videos when opposed to conventional content, especially regarding the center bias.

Keywords: Dataset, Salience, Unmanned Aerial Vehicles (UAV), Videos, Visual attention, eye tracking, surveillance.

1. Introduction

For a couple of decades now, we have witnessed the fast advances and growing use of UAVs for multiple critical applications. UAVs refer here to unmanned aerial vehicles, autonomous or monitored from remote sites. This imagery enables a broad range of applications from making vacation movies to drone races for mainstream civilian applications, from fire detection [1], wildlife counting [2] to journalism [3], precision agriculture and delivery services for professional applications, and from military aerial surveillance [4], drone-based warfare [5] to tracking moving targets [6], object, person or anomaly detection [7–9] for military applications.

The UAV imagery proposes a new representation of visual scenes that makes all these new applications possible. UAV vision is dominant and hegemonic [10]. The bird point of view modifies the perspective, size and features of objects [11]. Also, their high autonomy in conjunction with large-field of view camera permit to cover large areas in limited time duration. Besides, UAV sensors can be multi-modal and can include RGB, thermal, Infra-Red (IR), or multi-spectral sensors. Multiplying imagery modalities allows overcoming possible weaknesses of RGB-only [10]. For instance, occlusions

may be compensated by thermal information, and the capture of IR is desired for low-luminance environments [12].

UAV scene depiction is rich, comprehensive, and promising, which explains its success. But challenges to come are even more compelling. In [10], Edney-Browne wondered how the capacity of UAV capturing the external reality (visuality) is related to perceptual and cognitive vision in humans. Variations in UAV characteristics, such as perspective view and object size, may change viewers' attitudes towards content. Consequently, new visual attention processes may be triggered for this specific imaging. This means that studying UAV imagery in light of human visual attention not only opens the door to plenty of applications but also enables to gather further knowledge on perceptual vision and cognition.

Visual attention occurs to filter and sort out visual clues. Indeed, it is impossible to process all the information one perceives. Particular consideration should be dedicated to identifying which attentional processes are involved as they are diverse and aim at specific behaviors. For instance, one must make the distinction between overt and covert attention [13]. The former refers to a direct focus onto where points eyes and head. The latter relates more to the peripheral vision, where attention is directed without eye movements towards it. In practice, when an object of interest is detected in the area covered by the covert attention, one may make a saccade movement to direct the eyes from the overt area to this position. The context of visualization is also important. For instance, we make a distinction between two content exploration processes [14]: (1) A no constraint examination named free viewing. The observer is rather free from cognitive loads and is supposed to mainly use bottom-up or exogenous attention processes driven by external factors, e.g. content and environment stimuli. (2) A task-based visualization, such as surveillance for instance. Cognitive processes such as prior knowledge, willful plans, and current goals guide the viewer's attention. This is known as top-down or endogenous attention. A strict division is slightly inaccurate in that both top-down and bottom-up processes are triggered during a visual stimuli in a very intricate interaction [15]. Both processes are important and need to be studied through salience.

Visual salience means to represent attention in multimedia content as a probability distribution per pixels [16]. Salience analyses rest on the relation of visual attention to eye movements, and these latter are obtained through gaze collection with eye-trackers [17]. Saliency predictions help to understand computational cognitive neuroscience as it reveals attention behaviors such as center bias and spatial and temporal inhibition of return [15]. Multiple applications derive from saliency predictions such as compression [18], content-aware re-targeting, object segmentation [19], and detection [20,21].

Recently, there has been a growing interest on one particular application, which combines visual salience and UAV content. Information overload in the drone program and fatigue in military operators may have disastrous consequences for military applications [10]. New methods and approaches are required to detect anomaly in UAV footages and to ease the decision-making. Among them, we believe that computational models of visual attention could be used to simulate operators' behaviors [22]. Eventually, thanks to predictions, operators' workloads can be reduced by eliminating unnecessary footages segments.

To that end, it is necessary to develop new dynamic saliency models tailored to UAV content. The gain brought by deep-learning saliency models this last decade [18–21,23] has been more than significant. This improvement comes with the definition and the design of large-scale eye tracking datasets, from which a ground truth can be defined. However, in the context of UAV content, there are very few eye-tracking datasets. This is the reason why we propose and present in this paper a new large-scale eye-tracking dataset, freely downloadable from internet.

The paper is organized as follows. In section 2, we first justify and elaborate on the need for large-scale eye-tracking databases for UAV videos. Then, we introduce the entire process of dataset creation in section 3. It describes the content selection, the experiment set up, and the implementation of fixations, saccades, and saliency maps. Section 4 presents an in-depth analysis of the dataset. The study is two-fold: it explores what ground truth should be used for salience studies, and brings to light

the fading of conventional biases in visual salience for UAV stimuli. Finally, conclusions are provided in section 5.

2. Related Work

While it is now rather easy to find eye tracking data on typical images [24–33] or videos [34–38], and that there are many UAV content datasets [7,39–50], it turns out to be extremely difficult to find eye-tracking data on UAV content. Indeed, very few works are dealing with eye-tracking data related to UAV content. This is even truer when we consider dynamic salience, which refers to salience for video content.

To the best of our knowledge, *EyeTrackUAV1* dataset released in 2018 [11] is the only public dataset available for studying the visual deployment over UAV video. There exist another dataset *AVS1K* [51]. However, *AVS1K* is, to the present day, not publicly available. We thus focus here on *EyeTrackUAV1*, with the awareness that all points below but the last apply to *AVS1K*.

EyeTrackUAV1 consists in 19 sequences (1280x720 and 30 frame per second (fps)) extracted from the *UAV123* database [43]. The sequence selection relied on content characteristics, which are the diversity of environment, distance and angle to the scene, size of the principal object, and the presence of sky. Precise binocular gaze data (1000 Hz) of 14 observers were recorded under free viewing condition, for every content. Overall, the dataset comprises eye-tracking information on 26599 frames, which represents 887 seconds of video. In spite of a number of merits, this dataset presents several limitations for saliency prediction applications. These limitations have been listed in [23]. We briefly summarize them below:

- UAV may embed multi-modal sensors during the capture of scenes. Besides conventional RGB cameras, to name but a few thermal, multi-spectral, and infrared cameras consist of typical UAV sensors. Unfortunately, *EyeTrackUAV1* lacks non-natural content, which is of great interest for the dynamic field of salience. As already mentioned, combining content from various imagery in datasets is advantageous for numerous reasons. It is necessary to continue efforts toward the inclusion of more non-natural content in databases.
- In general, the inclusion of more participants in the collection of human gaze is encouraged. Indeed, reducing variable errors by including more participants in the eye tracking experiment is beneficial. It is especially true in the case of videos as salience is sparse due to the short displaying duration of a single frame. With regards to evaluation analyses, some metrics measuring similarity between saliency maps consider fixation locations for saliency comparison (e.g. any variant of Area Under the Curve (AUC), Normalized Scanpath Saliency (NSS), and Information Gain (IG)). Having more fixation points is more convenient for the use of such metrics.
- *EyeTrackUAV1* contains eye-tracking information recorded during free viewing sessions. That is, no specific task was assigned to observers. Several applications, for UAV and conventional imaging, could benefit from the analysis and reproduction of more top-down attention, related to a task at hand. More specifically, for UAV content, there is a need for specialized computational models for person detection or anomaly detection.
- Even though there are about 26599 frames in *EyeTrackUAV*, they come from "only" 19 videos. Consequently, this dataset just represents a snapshot of the reality. We aim to go further by introducing more UAV content.

To extend and complete the previous dataset and to tackle these limitations, we have created the *EyeTrackUAV2* dataset, introduced below.

3. *EyeTrackUAV2* dataset

This section introduces the new dataset *EyeTrackUAV2* aiming at tackling issues mentioned above. *EyeTrackUAV2* includes more video content than its predecessor *EyeTrackUAV1*. It involves more

participants, and considers both free and task-based viewing. In the following subsections, we first elaborate on the selection of video content, followed by a description of the eye-tracking experiment. It includes the presentation of the eye-tracking apparatus, the experiment procedure and setup, and the characterization of population samples. Finally, we describe the generation of the human ground truth, i.e. algorithms for fixation and saccade detection as well as saliency map computation.

3.1. Content selection

Before collecting eye-tracking information, experimental stimuli were selected from multiple UAV video datasets. We paid specific attention to select videos suitable for both free and task-based viewing as experimental conditions. Also, the set of selected videos has to cover multiple UAV flight altitudes, main surrounding environments, main sizes of observed objects and angles between the aerial vehicle and the scene, as well as the presence or not of sky. We consider these characteristics favor the construction of a representative dataset of typical UAV videos, as suggested in [11].

After examining a number of UAV datasets (UCF's dataset [1], VIRAT [39], MRP [40], the privacy-based mini-drones dataset [41], the aerial videos dataset described in [42], UAV123 [43], DTB70 [45], Okutama-Action [46], VisDrone [52], CARPK [47], SEAGULL [48], DroneFace [49], and the aerial video dataset described in [44]), a total of 43 videos (RGB, 30 fps, 1280x720 or 720x480) have been selected from 3 different databases, *VIRAT*, *UAV123* and *DTB70*. These three databases are exhibiting different content for various applications, which makes the final selection representative of the UAV ecosystem. Table 2 reports the number of sequences selected from each database and details their native resolution, duration and frame number. We present below the main characteristics of the three selected datasets:

- *UAV123* includes challenging UAV content annotated for object tracking. We restrict the content selection to the first set, which includes 103 sequences (1280x720 and 30 fps) captured by an off-the-shelf professional-grade UAV (DJI S1000) tracking various objects in a range of altitudes comprised between 5-25 meters. Sequences include a large variety of environments (e.g. urban landscapes, roads, and marina), objects (e.g. cars, boats, and persons) and activities (e.g. walking, biking, and swimming) as well as present many challenges for object tracking (e.g. long- and short-term occlusions, illumination variations, viewpoint change, background clutter, and camera motion).
- Aerial videos in the VIRAT dataset were manually selected (for smooth camera motion and good weather conditions) from rushes of a total amount of 4 hours in outdoor areas with broad coverage of realistic scenarios for real-world surveillance. Content includes "single person", "person and vehicle", and "person and facility" events, with changes in viewpoints, illumination, and visibility. The dataset comes with annotations of moving object tracks and event examples in sequences. These videos (720x480 and 30 fps) exhibit quite low quality and include content shot in infra-red.
- The 70 videos (RGB, 1280x720 and 30 fps) from DTB70 dataset are manually annotated with bounding boxes for tracked objects. Sequences were shot with a DJI Phantom 2 Vision+ drone or were collected from YouTube to add diversity in environments and target types (mostly humans, animals, and rigid objects). There is also a variety of camera movements (both translation and rotation), short- and long-term occlusions, and target deformability.

Table 1 presents the sequences which have been extracted from their original datasets. Video characteristics such as duration, spatial and temporal complexity are also given. Native resolutions are provided by original dataset in Table 2.

[1] http://crcv.ucf.edu/data/UCF_Aerial_Action.php

ID	Video	Dataset	Number of frames	Start frame	End frame	Duration (msec)	SI	TI	Altitude	Environment	Object size	Horizontal line (sea, sky)	Main angle
1	09152008flight2tape1_3 (crop 1)	VIRAT	120	1	120	4000	0,455	32	High	Urban military - **IR**	Small	False	Oblique
2	09152008flight2tape1_3 (crop 2)		367	137	503	12234	0,474	35	High	Urban military - **IR**	Small	False	Oblique
3	09152008flight2tape1_3 (crop 3)		3178	4735	7912	105934	0,452	43	Intermediate	Urban military	Medium, Small	False	Oblique
4	09152008flight2tape1_5 (crop 1)		972	218	1189	32400	0,467	37	Intermediate	Urban military	Medium, Small	False	Oblique
5	09152008flight2tape1_5 (crop 2)		1715	4555	6269	57167	0,461	45	Intermediate	Urban military	Medium, Small	False	Oblique
6	09152008flight2tape2_1 (crop 1)		1321	1	1321	44034	0,484	40	Intermediate, Low	Urban military	Medium, Big	False	Oblique
7	09152008flight2tape2_1 (crop 2)		1754	2587	4340	58467	0,484	41	High	Roads rural - **IR**	Small	False	Oblique
8	09152008flight2tape2_1 (crop 3)		951	4366	5316	31700	0,482	33	Intermediate	Urban military	Medium, Big	False	Oblique
9	09152008flight2tape2_1 (crop 4)		1671	6482	8152	55700	0,452	32	High	Roads rural	Medium	False	Oblique, Vertical
10	09152008flight2tape3_3 (crop 1)		2492	3067	5558	83067	0,474	42	Intermediate	Urban military	Small	False	Oblique
11	09162008flight1tape1_1 (crop 1)		1894	1097	2990	63134	0,448	39	Low	Urban military, Roads rural	Medium, Small	False	Oblique
12	09162008flight1tape1_1 (crop 2)		1416	4306	5721	47200	0,477	29	Intermediate, High	Urban military	Small	False	Oblique
13	bike2	UAV123	553	1	553	18434	0,468	22	Intermediate	Urban, building	Small, Very small	True	Horizontal
14	bike3		433	1	433	14434	0,462	19	Intermediate	Urban, building	Small	True	Horizontal
15	building1		469	1	469	15634	0,454	12	Intermediate	Urban, building	Very Small	True	Horizontal
16	building2		577	1	577	19234	0,471	37	Intermediate	Urban, building	Medium, Small	True	Horizontal
17	building3		829	1	829	27634	0,451	27	High	Urban in desert	Small	True	Horizontal
18	building4		787	1	787	26234	0,464	29	High, Intermediate	Urban in desert	None	True, False	Horizontal, Oblique
19	car1		2629	1	2629	87634	0,471	59	Low, Intermediate	Road rural	Big, Medium	True	Oblique
20	car11		337	1	337	11234	0,467	31	High	Suburban	Small	True, False	Horizontal, Oblique
21	car12		499	1	499	16634	0,467	39	Low	Road urban, sea	Medium, Small	True	Horizontal
22	car13		415	1	415	13834	0,461	26	High	Urban	Very very small	False	Oblique, Vertical
23	car14		1327	1	1327	44234	0,471	25	Low	Road suburban	Medium	False	Oblique
24	car15		469	1	469	15634	0,471	18	Intermediate	Road towards urban	Small, Very small	True	Oblique
25	car2		1321	1	1321	44034	0,464	24	Intermediate	Road rural	Medium	False	Oblique, Vertical
26	car3		1717	1	1717	57234	0,467	27	Intermediate	Road rural	Medium	False	Oblique, Vertical
27	car4		1345	1	1345	44834	0,462	23	Intermediate, Low	Road rural	Big	False	Oblique, Vertical
28	car7		1033	1	1033	34434	0,464	18	Intermediate	Road suburban	Medium	False	Oblique
29	car9		1879	1	1879	62634	0,470	23	Intermediate, Low	Road suburban	Medium	False, True	Oblique, Horizontal
30	person22		199	1	199	6634	0,456	31	Low	Urban sea	Medium, Big	True	Horizontal
31	truck2		601	1	601	20034	0,453	24	High	Urban road	Small	True	Horizontal
32	truck3		535	1	535	17834	0,472	18	Intermediate	Road towards urban	Small, Very small	True	Oblique
33	truck4		1261	1	1261	42034	0,466	17	Intermediate	Road towards urban	Small	True	Oblique, Horizontal
34	wakeboard8		1543	1	1543	51434	0,472	39	Low	Sea urban	Medium, Big	True, False	Oblique, Vertical, Horizontal
35	Basketball	DTB70	427	1	427	14234	0,477	48	Intermediate	Field suburban	Medium	True	Oblique
36	Girl1		218	1	218	7267	0,481	31	Low	Field suburban	Big	True	Horizontal
37	Girl2		626	1	626	20867	0,482	30	Low	Field suburban	Big	True	Horizontal
38	ManRunning1		619	1	619	20634	0,483	23	Low	Field suburban	Big	True	Horizontal, Oblique
39	ManRunning2		260	1	260	8667	0,484	27	Low	Field suburban	Very big	False	Vertical, Oblique
40	Soccer1		613	1	613	20434	0,476	57	Low, Intermediate	Field suburban	Very big, Big	True	Horizontal
41	Soccer2		233	1	233	7767	0,475	24	High	Field suburban	Small	True	Oblique
42	StreetBasketball1		241	1	241	8034	0,379	37	Low	Field urban	Big	True, False	Oblique, Vertical
43	Walking		395	1	395	13167	0,476	31	Low	Field suburban	Big, Very big	True	Oblique

Table 1. Stimuli ID and name, their original dataset, number of frames together with starting and ending frame number, duration and native resolution.

Dataset	Native resolution	Proportion of content seen per degree of visual angle (%)	Videos number	Frames number (30 fps)	Duration (sec)
VIRAT [39]	720 x 480	1,19	12	17851	595,03
UAV123 [43]	1280 x 720	0,44	22	20758	691,93
DTB70 [45]	1280 x 720	0,44	9	3632	121,07
Overall			**43**	**42241**	**1408,03 (23:28 min)**

Table 2. Stimuli original datasets.

	Number of frames				Duration (MM:SS)			
	VIRAT	UAV123	DTB70	Overall	VIRAT	UAV123	DTB70	Overall
Total	17851	20758	3632	42241	09:55	11:32	02:01	23:28
Average	1488	944	404	982	00:50	00:31	00:13	00:33
Standard Deviation	847	615	177	727	00:28	00:21	00:06	00:24
Minimum	120	199	218	120	00:04	00:07	00:07	00:04
Maximum	3178	2629	626	3178	01:46	01:28	00:21	01:46

Table 3. Basic statistics on selected videos.

3.2. Content Diversity

To present the diversity of selected UAV sequences, Figure 1 illustrates the first frame of every content. Visual stimuli cover a variety of visual scenes in different environments (e.g. public and military environments, roads, buildings, sports, and port areas, etc.) and different moving or fixed objects (e.g. people, groups of people, cars, boats, bikes, motorbikes, etc.). Selected videos were captured from various flight heights and different angles between the UAV and the ground (allowing or not the presence of sky during their observation). This information is reported per sequence in Table 1. Additionally, we considered various video duration as the length of the video may possibly impact the behavior of observers due to fatigue, resulting in a lack of attention and more blinking artifacts [10,53].

To quantitatively show the diversity of selected videos, we have computed temporal and spatial complexity [54], named TI ($\in [0, +\infty[$) and SI ($\in [0, +\infty[$), respectively. These features are commonly used in image quality domain for characterizing the properties of selected images. They characterize the maximum standard deviation of spatial and temporal discrepancies over the entire sequence. The higher a measure is, the more complex is the content. TI and SI are reported per sequence in Table 1. The range of temporal complexity in sequences is broad, displaying the variety of movements present in sequences. Spatial measures are more homogeneous. Indeed, the spatial complexity is due to the bird point of view of the sensor. The aircraft high up position offers access to a large amount of information.

Three sequences, extracted from the VIRAT dataset, were captured by IR cameras. As a side note, finding non-natural content for UAV of sufficient quality in publicly available datasets was difficult.

Finally, Table 3 presents basic statistics of the database in terms of number of frames and duration. The 43 selected videos are now referred to as test stimuli in the following section, which presents the experiment design.

3.3. Experimental design

To record the gaze deployment of subjects while viewing UAV video sequences displayed onscreen, it is required to define an experimental methodology. All the details are presented below.

3.3.1. Eye-tracking apparatus

A specific setup is designed to capture eye-tracking information on video stimuli. It includes a rendering monitor, an eye-tracking system, a control operating system, and a controlled laboratory test room.

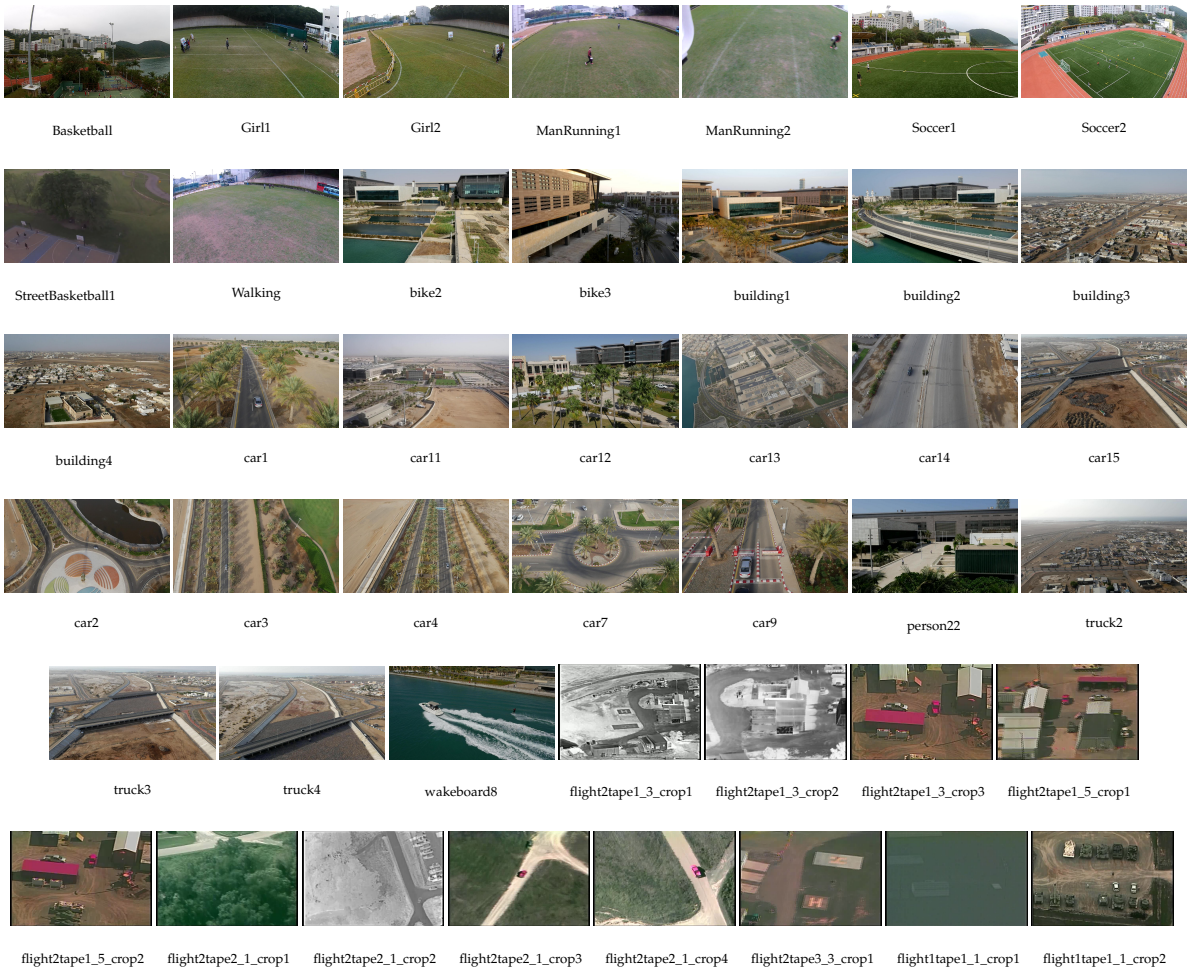


Figure 1. *EyeTrackUAV2* dataset: first frame of each sequence.

To run the experiment and collect gaze information, we used the EyeLink® 1000 Plus eye-tracking system, in the head free-to-move *remote mode*, taking advantage of its embedded 25mm camera lens. The eye tracker principle is to detect and record the IR illuminator reflection rays on the observer's pupil. This system enables the collection of highly precise gaze data at a temporal frequency of 1000 Hz and a spatial accuracy between the visual angle range of 0.25 and 0.50 degree, according to the manufacturer. To prevent errors in the robust algorithm for the detection of observers' pupils [53], participants were asked to remove any excess of mascara if need be. Also, the eye tracker's camera was configured for each subject, without affecting the corresponding distance between them. This configuration guarantees to achieve an optimal detection of the observer's eyes and head sticker.

The experimental monitor which displayed stimuli was a 23.8 inches (52,70 x 29,65 cm) DELL P2417H computer monitor display [2] with full HD resolution (1920x1080) at 60 Hz and with a response time of 6 ms. As suggested by both the International Telecommunication Union (ITU)-Broadcasting service (Television) (BT).710 [55] and manufacturer, observers sited in distance of about 3H (1m $\pm$ 10cm) from the monitor, where H corresponds to the stimuli display height so that observers have an assumed spatial visual angle acuity of one degree. Moreover, the eye tracker camera was placed 43 cm away from the experimental display, and thus about 67 cm from participants. Based on this setting,

[2] https://www.dell.com/cd/business/p/dell-p2417h-monitor/pd

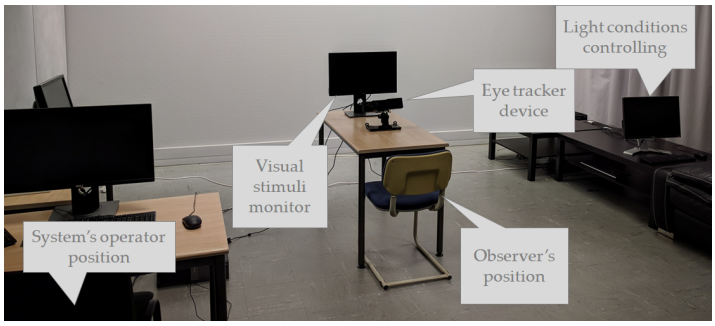


Figure 2. Experiment setup.

there are 64 pixels per degree of visual angle in each dimension, and the display resolution is about 30x17 visual degrees.

Regarding software, the MPC-HC video player [3], considered as one of the most lightweight open-source video players, rendered the experimental video stimuli. Also, we took advantage of the Eyelink toolbox [56] as it is part of the 3rd version of Psychophysics Toolbox Psychtoolbox-3 (PTB-3) [4] and added in-house communication processes [5] for sync between control and display systems. The control system consists of an additional computer, used by the experimenter to configure and control the eye-tracking system with an Ethernet connection.

Eventually, eye-tracking tests were performed in a room with controlled constant light conditions. The performed calibration set the constant ambient light conditions at approximately 36.5 cd/m^2, i.e. 15% of the maximum stimuli monitor brightness - 249 cm/m^2 - as recommended by the ITU-BT.500 [57], with the i1 Display Pro X-Rite® system.

Figure 2 illustrates the experimental setup used during the collection of gaze information. We can observe the arrangement of all the systems described above.

3.3.2. Stimuli presentation

The random presentation of stimuli in their native resolution centered on the screen prevents ordering, resizing, and locating biases. Knowing that the monitor resolution is higher than that of selected sequences, video stimuli were padded with mid-grey. Additionally, to avoid possible biases in gaze allocation, a 2-second sequence of mid-gray frames was presented before playing a test sequence. Please note that the amount of original information contained in a degree of visual angle is not the same for VIRAT sequences than for other database content, as specified in Table 2.

Before starting the experiment, a training session is organized to get the subject familiar with the experiment design. It includes a calibration procedure and its validation followed by the visualization of one video. This UAV video, the sequence *car4* from the DTB70 dataset, is additional to test stimuli to avoid any memory bias. Once subjects completed the training session, they could ask questions to experimenters before taking part into test sessions.

Regarding test sessions, they start with calibration and its validation before the visualization of 9 videos, during which subjects do or do not perform a task. To ensure the optimal quality of the collected gaze data, each participant took part in five test sessions. Splitting the experiment into sessions decreases the tiredness and lack of attention in observers. Also, this design enables frequent calibration so that recordings do not suffer from the decrease of accuracy in gaze recordings with time [53].

---

[3] https://mpc-hc.org/
[4] http://psychtoolbox.org/
[5] LS2N, University of Nantes

With regards to calibration, the eye-tracking system is calibrated for each participant, following a typical 13 fixed-point detection procedure [53]. Actually, experimenters started tests with a 9-point strategy for calibration (subject 1 to 17 in Free Viewing (FV)) but realized that gaze collection is more accurate with a 13-point calibration. The calibration reaches validation when the overall deviation of both eye positions is approximately below the fovea vision accuracy (e.g. a degree of visual angle [53,58]). The calibration procedure is repeated until validation.

The participation of an observer in the experiment lasts about 50 minutes. It includes test explanations, forms signing, and taking part in the training and the five test sessions. This duration is acceptable regarding the number of sessions and the fatigue in subjects.

3.3.3. Visual tasks to perform

EyeTrackUAV2 aims to investigate two visual tasks. Indeed, we want to be able to witness visual attention processes triggered by top-down (or goal-directed) and bottom-up (or stimulus-driven) attention. Accordingly, we defined two visual tasks participants have to perform: the first condition is a Free Viewing (FV) task while the second relates to a surveillance-viewing Task (Task). The former task is rather common in eye-tracking tests [24,31,33,38,59,60]. Observers were simply asked to observe visual video stimuli without performing any task. For the surveillance-viewing task, participants were required to watch video stimuli and to push a specific button on a keyboard each time they observe a new - meaning not presented before - moving object (e.g. people, vehicle, bike, etc.) in the video. The purpose of this task is to simulate one of the basic surveillance procedures in which targets could be located anywhere when the visual search process was performed [61]. After reviewing typical surveillance systems' abilities [62], we have decided to define our task as object detection. The defined object detection task is compelling in that it encompasses target-specific training (repeated discrimination of targets and non-targets) and visual search scanning (targets potentially located anywhere) [61]. The only task-related behavior not explored within this task is cue training (targets likely to be co-located with more salient objects or events). The surveillance-viewing task is especially interesting for a military context, in which operators have to detect discrepancies in drone videos.

3.3.4. Population

Overall, 30 observers participated in each phase of the test. Tested population samples were different for these two viewing conditions. They were carefully selected to be as representative as possible of the entire population. For instance, they include people from more than 12 different countries, namely Algeria (3%), Brazil, Burundi, China, Colombia (10%), France (67 %), Gabon, Guinea, South Arabia, Spain, Tunisia, and Ukraine. Additionally, we achieved gender and almost eye-dominance balance in both phases tests. Table 4 presents the detailed population characteristics for both tasks.

Each observer has been tested for visual acuity and color vision with Ishihara and Snellen tests [63,64]. Any failure to these tests motivated the dismissal of the person from the experiment. Before running the test, the experimenter provided subjects with written consents and information forms, together with oral instructions. This process made sure of the consent of participants and their understanding of the experiment process. It also ensures an anonymous data collection.

Sample statistics	FV	Task	Total
Participants	30	30	60
Females	16	16	32
Males	14	14	28
Average age	31,7	27,9	29,8
Std age	11,0	8,5	10,0
Min age	20	19	19
Max age	59	55	59
Left dominant eye	19	9	28
Right dominant eye	11	21	32
Participants with glasses	0	4	4

Table 4. Population characteristics.

Figure 3. Stimulus displayed in its native resolution, and padded with mid-gray to be centered. Colored information relates to Equation 1.

3.4. Post-processing of eye-tracking data

First, the conversion of collected raw signals into the pixel coordinate system of the original sequence leads to what we refer to as binocular gaze data. Let us precise that the origin of coordinates is the top-left corner. Then, any gaze coordinates out of range are evicted, as they do not represent visual attention on stimuli. Once transformed and filtered, we extract fixation and saccade information and create saliency maps (also called grayscale heatmaps) from gaze data. The remainder of this section describes all post-processing functions.

3.4.1. Raw data

At first, coordinates of the collected binocular gaze data were transformed into the pixel coordinate system of the visual stimulus. Additionally, we addressed the original resolution of sequences. Coordinates outside the boundaries of the original resolution of the stimulus were filtered out as they were not located in the video stimuli display area. The following formula presents how the collected coordinates are transformed for both eyes:

$$\begin{cases} x_S &= \lfloor x_D - \frac{R^X_D - R^X_S}{2} \rfloor \\ y_S &= \lfloor y_D - \frac{R^Y_D - R^Y_S}{2} \rfloor \end{cases} \quad (1)$$

where, (x_S, y_S) and (x_D, y_D) are the spatial coordinates on the stimulus and on the display, respectively. The operator $\lfloor . \rfloor$ allows to keep the coordinates if the coordinates are within the frame of the stimulus. Otherwise, the coordinate is discarded. (R^X_S, R^Y_S) and (R^X_D, R^Y_D) represent the stimulus resolution and the display resolution, respectively. For more clarity, Figure 3 displays the terms of the equation. Once this remapping has been done for both eyes, the spatial binocular coordinates is simply given by the average of the spatial coordinates of left and right eyes.

For the surveillance-viewing task, each gaze point was assigned with the relative information of the button reaction. We denote in raw data a button activation (respectively no button reaction) with the Boolean value 1 (respectively 0). Finally, for convenience, we have sorted the positions of the observer's dominant eyes and included them in raw gaze data.

3.4.2. Fixation and saccade event detection

To retrieve fixations from eye positions, we used the Dispersion-Threshold Identification (I-DT) [65] from the EyeMMV and LandRate toolboxes [66,67]. This algorithm performs "two-step" spatial and temporal thresholds. As exposed in [67,68], thanks to the very high precision of our eye-tracking equipment, we can combine the two-step spatial thresholds in one operation, as both thresholds have the same value. Ultimately, in our context, this algorithm conceptually implements a spatial noise removal filter and a temporal threshold indicating the minimum fixation duration. We have selected

the minimum threshold values from the state of the art to ensure the performance of the fixation detection algorithm. Accordingly, spatial and temporal thresholds were selected to be equal to 0.7 degree of the visual angle and 80 ms [69], respectively. Finally, saccade events were calculated based on the computed fixations considering that a saccade corresponds to eye movements between two successive fixation points.

When considering raw data of the dominant eye, I-DT exhibits a total number of fixations of 1 239 157 in FV and 1 269 433 in Task.

3.4.3. Saliency maps

Saliency maps are a 2D topographic representation indicating the ability of an area to attract observers' attention. It is common to represent the salience of an image thanks to either its saliency map or by its colored representation, called heatmap. Saliency maps are usually computed by convolving the fixation map, gathering observers' fixations, with a Gaussian kernel representing the foveal part of our retina. More details can be found in [60].

For video sequence, there is one saliency map for each frame of test sequences. It could be debatable to do so for temporal analyses, but it is current practices to deal with videos as a succession of frames in visual media processing (e.g. compression, High Dynamic Range (HDR) video tone mapping, and dynamic saliency).

We took benefit from the high frequency of acquisition of the eye-tracker system to compute saliency maps directly from raw gaze data (in pixel coordinates). Thus, our saliency maps are free from any biases that could be introduced by any fixation extractor algorithms. Hence, the generated saliency maps include fixation and saccade information, without distinction.

To indicate salient regions of each frame, we followed the method described in [66], with parameters derived from the experimental setup (e.g., a grid size of a pixel, a standard deviation of 0.5 degree of angle i.e. σ =32 pixels, and a kernel size of 6σ). For visualization purposes, heat maps were normalized between 0 and 255.

Figure 4 presents saliency maps obtained for both attention conditions in frame 100 of seven sequences. We have selected frame 100 to get free from the initial center-bias in video exploration occurring during the first seconds. These examples illustrate, for instance, the sparsity of salience in videos in free viewing, while task-based attention usually presents more salient points, more dispersed in the content than FV, depending on the task and attention-grabbing objects.

3.5. EyeTrackUAV2 in brief

We have created a dataset containing binocular gaze information collected during two viewing conditions (free viewing and task) over 43 UAV videos (30 fps, 1280x720 and 720x480 - 42241 frames, 1408 seconds) observed by 30 participants per condition, leading to 1 239 157 fixations in free viewing and 1 269 433 in task-viewing for dominant eyes positions. Notably, selected UAV videos sowing diversity in rendered environments, movement and size of objects, aircraft flight heights and angles to the ground, duration, size, and quality. This dataset overcomes the limitations of *EyeTrackUAV1* in that it enables investigations of salience in more test sequences, on larger population samples, and for both free-viewing and task-based attention. Additionally, and even though they are still too few, three IR videos are part of visual stimuli.

Fixations, saccades, and saliency maps were computed - for both eyes in additive and averaged fashions (see Binocular and BothEyes scenarios described later) and for the dominant eye - and are publicly available with original content and raw data on our FTP [6]. The code in MATLAB to generate all ground truth information is also made available.

[6] ftp.ivc.polytech.univ-nantes.fr

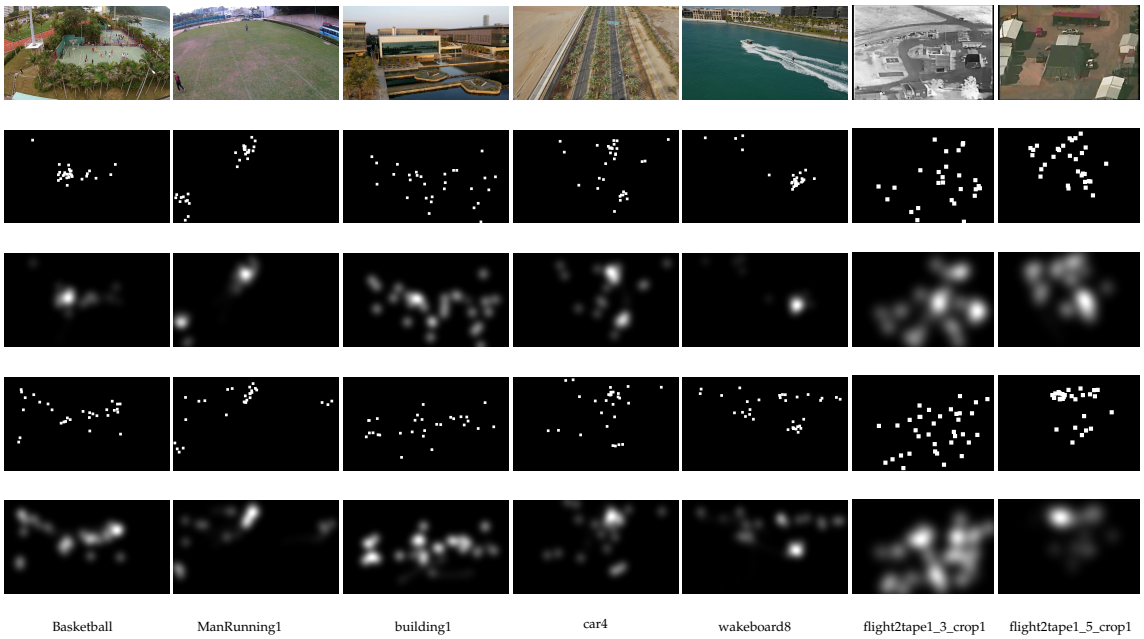


Figure 4. Frame 100 of seven sequences of *EyeTrackUAV2* dataset, together with saliency and fixation maps generated based on gaze data of dominant eye. Results are presented for both types of attention. The first row presents sequences hundredth frame, the second fixations for FV, the third saliency maps for FV, the fourth fixations for task, and the fifth saliency maps for Task.

4. Analyses

In this section, we characterize the proposed *EyeTrackUAV2* database. On one hand, we compare eye positions and salience between the different attention conditions. Such information may be of great importance to generate the ground truth on a case-by-case basis. On the other hand, UAV videos induce new visual experiences. Consequently, observers exhibit different behaviors towards this type of stimuli. Therefore, we investigate whether the center bias, one of the main viewing tendencies [70], still applies to *EyeTrackUAV2* content.

4.1. Six different ground truths

The first question we address concerns the method used to determine the ground truth. In a number of papers, researchers use the ocular dominance theory in order to generate a ground truth. This theory relies on the fact that the human visual system favors the input of one eye over the other should binocular images be too disparate on the retinas. However, the cyclopean theory gains more and more momentum [71,72]. It alleges that vision processes approximate a central point between two eyes, from which an object is perceived. Furthermore, lately, manufacturers achieved major improvements in eye-tracking systems. They are now able to record and calibrate the positions of both eyes separately. This allows for exploring what are the best practices to create salience ground truth [71–73].

We therefore propose to evaluate the potential errors made when different methods for creating the ground truth are used. We tested six methods, namely Left (L), Right (R), Binocular (B), Dominant (D), non Dominant (nD), Both Eyes (BE), under the two visual attention conditions, Free Viewing (FV) and surveillance-viewing Task (Task). B corresponds to the average position between the left and right eyes and can be called version signal. BE includes the positions of both L and R positions, and hence comprises twice more information than other scenarios. nD has been added to estimate the gain made when using dominant eye information. Estimating the relevance of the aforementioned methods will help to decide which ground truth scenario should be used depending on the precision and accuracy

		FV						Task					
		Average		max		min		Average		max		min	
		hor x	vert y	hor x	vert y	hor x	vert y	hor x	vert y	hor x	vert y	hor x	vert y
Binocular	EyeNonDom	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Binocular	Right	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Binocular	Left	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Binocular	Dominant	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
EyeNonDom	Right	**0,16**	**0,17**	**0,21**	**0,24**	**0,12**	**0,10**	0,31	0,28	0,37	0,34	0,25	0,25
EyeNonDom	Left	0,31	0,34	0,40	0,44	0,22	0,27	**0,15**	**0,12**	**0,20**	**0,15**	**0,11**	**0,10**
EyeNonDom	Dominant	**0,46**	**0,51**	**0,57**	**0,63**	**0,39**	**0,42**	**0,46**	**0,40**	**0,56**	**0,45**	**0,39**	**0,36**
Right	Left	**0,46**	**0,51**	**0,57**	**0,63**	**0,39**	**0,42**	**0,46**	**0,40**	**0,56**	**0,45**	**0,39**	**0,36**
Right	Dominant	0,31	0,34	0,40	0,44	0,22	0,27	**0,15**	**0,12**	**0,20**	**0,15**	**0,11**	**0,10**
Left	Dominant	**0,16**	**0,17**	**0,21**	**0,24**	**0,12**	**0,10**	0,31	0,28	0,37	0,34	0,25	0,25
Binocular	BothEyes	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
EyeNonDom	BothEyes	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Right	BothEyes	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Left	BothEyes	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18
Dominant	BothEyes	0,23	0,26	0,29	0,31	0,19	0,21	0,23	0,20	0,28	0,23	0,20	0,18

Table 5. MAE in eye positions depending on scenarios for Free Viewing and Task viewing, in degree per pixels. Results are provided in average, minimum and maximum over sequences and observers. Highest values are emphasised in red, the least in blue.

one requires [73]. We performed two evaluations: the first directly on eye positions and the second on human saliency maps.

4.1.1. Mean of Absolute Error of eye positions

To characterize the error made when choosing a scenario over another, we compute the Mean of Absolute Error (MAE) between eye positions for both viewing conditions:

$$MAE_x^{(i,j)} = \frac{1}{N}\sum_n |x_{s,n}^{M_i} - x_{s,n}^{M_j}|$$
$$MAE_y^{(i,j)} = \frac{1}{N}\sum_n |y_{s,n}^{M_i} - y_{s,n}^{M_j}|$$

With $MAE_x^{(i,j)}$ and $MAE_y^{(i,j)}$ the Mean Absolute Error for x and y axis, respectively, and for methods M_i and M_j. Method refers here to Left (L), Right (R), Binocular (B), Dominant (D), non Dominant (nD), Both Eyes (BE). N denotes the number of gaze samples for one sequence, for all observers. Note that $MAE_x^{(i,i)} = MAE_y^{(i,i)} = 0$.

Table 5 reports the average, maximum and minimum MAE over sequences for all comparisons of scenarios. We observe several interesting behaviors, such as the decrease of error when using the dominant eye in a single-eyed method, or the fixed error made when using both eyes positions. But what really is noteworthy is that the maximum average MAE is about half of a degree of visual angle, the maximum error value being 0,63.

Half a degree of visual angle is usually the least value of Gaussian kernel used to filter eye positions (or fixations) when creating saliency maps. We thus extend the analysis with a comparison of methods in terms of similarity between saliency maps. We want to know if it makes a difference to use different eye positions to generate salience ground truth. Indeed, this would help the scientific community to know how to use eye-tracking data for saliency depending on which information is available. It is necessary to apprehend, when downloading a dataset, what is the possible error made by using only the dominant eye for instance.

4.1.2. Similarity of human saliency maps

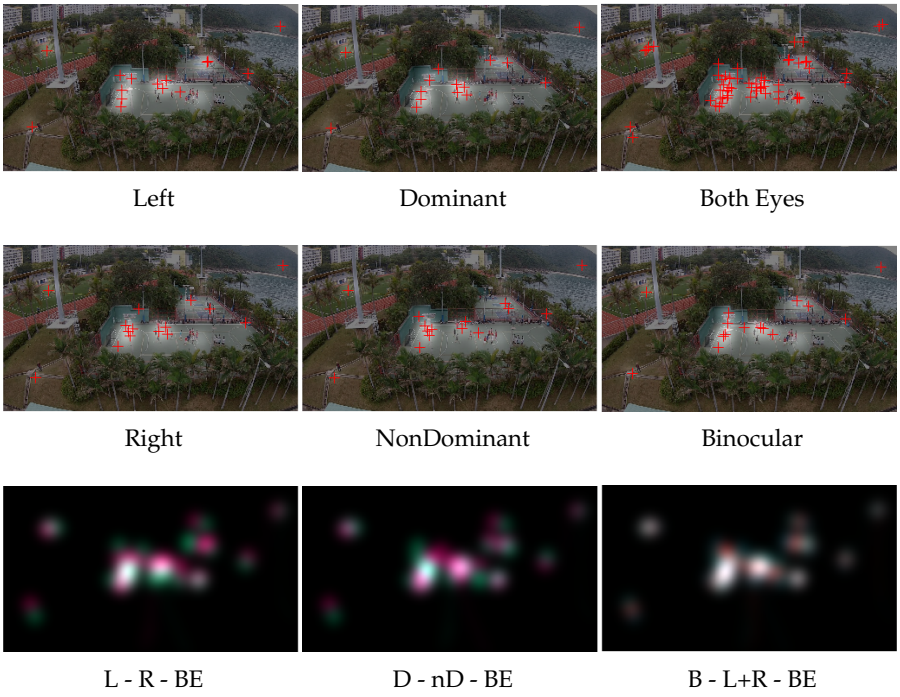


Figure 5. Qualitative comparison of saliency maps for all scenarios on Basketball, frame 401. Saliency and fixations are displayed in transparency over the content. The last row compares scenarios: first scenario is attributed to the red channel, the second to green and the last to blue. When fully overlapping, the pixel turns white.

Generating saliency maps implies Gaussian filtering with a rather large kernel when compared to MAE values.For instance, the Gaussian kernel we used here is 3°. We thus question whether selecting a ground truth scenario over another makes a significant difference for saliency studies.

Thus, we decided to compare saliency maps generated for the six scenarios defined above. Illustrations of scenarios saliency maps and fixations as well as methods comparisons are presented in Figure 5. Below is presented the quantitative evaluation.

We run a cross-comparison on six well-used quality metrics: Correlation Coefficient (CC) ($\in [-1,1]$), Similarity (SIM) ($\in [0,1]$) the intersection between histograms of saliency, AUC Judd and Borji ($\in [0,1]$), NSS ($\in]-\infty,+\infty[$), and IG ($\in [0,+\infty[$), which measures on average the gain in information contained in the saliency map compared to a prior baseline ($\in [0,+\infty[$). We did not report Kullback Leibler divergence (KL) ($\in [0,+\infty[$) as we favored symmetric metrics. Moreover, even though symmetric in absolute value, IG provides different scores depending on fixation maps. We thus compared scenarios for fixation maps of both methods, which leads to two IG measures. More details on metrics and metrics behaviors are given in [60,74,75].

To verify if scenarios are different, we have conducted ANOVA and multi-comparison analyses on the scores obtained by measures. All metrics show statistically different results ($p << 0.001$) except for AUC Borji, AUC Judd, and NSS. Thus, the analysis discards those metrics. However, it shows that for those metrics using a scenario over another makes no significant difference.

Table 6 presents the results of measures that present a significant difference when comparing saliency maps of two scenarios. We confirm here the results hinted by MAE scores:

- There is a high similarity between scenarios saliency maps. As expected, scores are pretty high (or low for IG), which indicates the high similarity between scenarios.

		FV				Task			
		CC ↑	SIM ↑	IG ↓		CC ↑	SIM ↑	IG ↓	
SM1	SM2			SM1-Fix1-SM2	SM2-Fix2-SM1			SM1-Fix1-SM2	SM2-Fix2-SM1
Binocular	Dominant	0,94	0,83	0,377	0,300	0,952	0,850	0,276	0,148
Binocular	EyeNonDom	0,95	0,84	0,370	0,301	0,952	0,849	0,283	0,163
Binocular	Left	0,94	0,83	0,371	0,301	0,948	0,843	0,264	0,192
Binocular	Right	0,94	0,83	0,390	0,324	0,944	0,838	0,304	0,152
Binocular	BothEyes	**0,98**	**0,90**	0,246	**0,139**	**0,983**	**0,916**	0,177	**0,012**
Dominant	BothEyes	0,96	0,87	**0,158**	0,374	0,967	0,873	**0,143**	0,248
EyeNonDom	BothEyes	0,97	0,87	0,167	0,394	0,966	0,872	0,144	0,228
Left	BothEyes	0,96	0,86	0,166	0,387	0,960	0,861	0,174	0,232
Right	BothEyes	0,96	0,86	0,181	0,416	0,960	0,862	0,147	0,279
Dominant	EyeNonDom	0,87	0,74	1,115	1,069	0,873	0,747	0,743	0,781
Dominant	Left	0,95	0,88	0,341	0,339	0,903	0,792	0,520	0,582
Dominant	Right	0,91	0,79	0,810	0,757	0,957	0,884	0,256	0,233
EyeNonDom	Right	0,96	0,88	0,346	0,342	0,902	0,793	0,587	0,519
Left	EyeNonDom	0,91	0,79	0,792	0,754	0,957	0,884	0,256	0,231
Left	Right	**0,85**	**0,72**	**1,176**	**1,121**	**0,850**	**0,725**	**0,877**	**0,782**
Mean		0,937	0,832	0,467	0,488	0,938	0,839	0,343	0,319
Std		0,037	0,052	0,340	0,295	0,038	0,053	0,230	0,234

Table 6. CC, SIM and IG results for scenarios cross-comparison. Red indicates the best scores, blue the least.

- Two-eyes-based saliency maps reach the best results. All metrics show the best results for comparisons including Binocular and BothEyes scenarios, the highest being the Binocular-BothEyes comparison.
- Left-Right and Dominant-NonDominant comparisons achieve worst results.
- It is possible to know the population main dominant eye through scenarios comparisons (not including two eyes information). When describing the population, we have seen that a majority of left-dominant-eye subjects participated in the FV test, while the reverse happened for the Task experiment. This fact is easy to notice in metric scores.

ANOVA and multi-comparison analyses characterize the differences between scenarios in Figure 6. We can see where stands the mean and standard deviation of CC scores for each scenario over the entire dataset. Scenarios having non-overlapping confidence intervals are statistically different.

We can see that for this metric, and this result applies to SIM and IG, it is recommended to use both eyes information, BothEyes and Binocular, as significantly higher similarity is reached for these two scenarios. The worst scenario, significantly in Task, is to favor the position of one eye over the other. Also, scenarios based on the dominant eye are obviously biased towards one eye, thus generating more errors than two-eyes but less than one-eye scenarios.

Overall, over six metrics, three do not find significant differences between the scenarios' saliency maps. The three others do and indicate that using both eye information must be favored. **Accordingly, the cyclopean theory takes a slight precedence over the ocular dominance theory in salience**. Moreover, it is recommended to favor datasets that record both eyes, and if not possible these that collect the dominant eye positions.

4.2. Biases in UAV videos

In conventional imaging, the center position is the best location to have access to most visual information of a content [70]. This fact leads to a well-known bias in visual attention named central bias. This effect may be associated with various causes. For instance, Tseng et al. [76] showed a contribution of photographer bias, viewing strategy, and to a lesser extent, motor, re-centering, and screen center biases to the center bias. They are briefly described below:

- The **photographer bias** often emphasizes objects in the content center through composition and artistic intent [76].
- Directly related to photographer bias, observers tend to learn the probability of finding salient objects at the content center. We refer to this behavior as a **viewing strategy**.

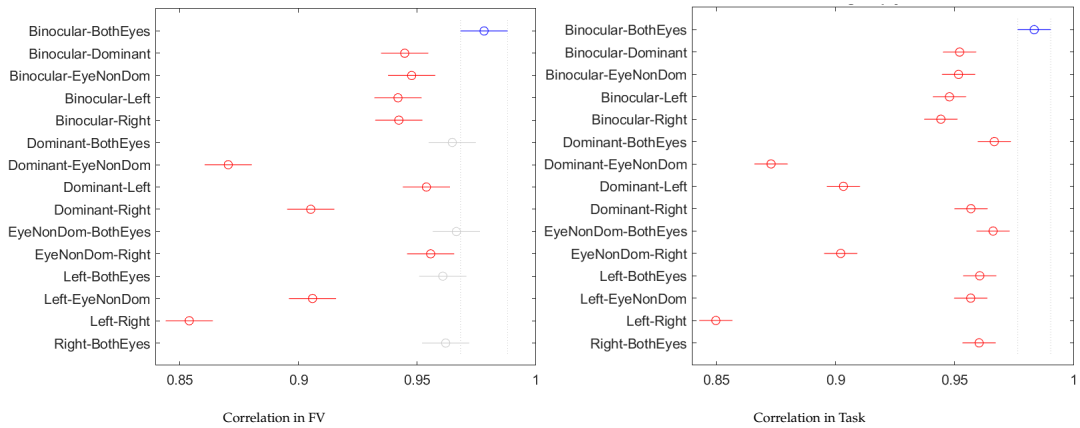


Figure 6. Muti-comparison on scenarios correlation measure.

- With regards to the Human Visual System (HVS), the central orbital position, that is when looking straight ahead, is the most comfortable eye position [77], leading to a **recentering bias**.
- Additionally, there is a **motor bias**, in which one prefers making short saccades and horizontal displacements [59,78].
- Lastly, onscreen presentation of visual content pushes observers to stare at the center of the screen frame [70]. This experimental bias is named the **screen center** bias.

The central bias is so critical in the computational modelling of visual attention that saliency models include this bias as prior knowledge or use it as a baseline to which saliency predictions are being compared. The center bias is often represented by a centered isotropic Gaussian stretched to the video frame aspect ratio [25,75].

The presence of this bias in UAV videos has already been questioned in [23]. In [23], authors showed that saliency models that heavily rely on the center bias were less efficient on UAV videos than on conventional video sequences. Therefore, we believe that the central bias could be less significant in drone footage as a result of the lack of photographer bias or due to UAV content characteristics. These latter comprise, but are not restricted to, the camera bird-point-of-view that changes objects semantic and size [11], the loss of pictorial depth cues [79] such as horizontal line [80], and the presence of camera movements [11]. To make this point clear, we propose to evaluate qualitatively and quantitatively the center bias for UAV videos.

4.2.1. Qualitative evaluation of biases in UAV videos

We evaluate the viewing tendency of observer thanks to the average saliency map, computed over the entire sequence. It is representative of the average position of gaze throughout the video sequence. It is used to observe potential overall biases, as it could be the case with the center bias. Figures 7 and 8 show the average saliency map for all sequences of *EyeTrackUAV2* dataset, generated from D scenario, for both free-viewing and task-viewing conditions. Several observations can be made.

Content-dependent center bias. We verify here the content-dependence of the center bias in UAV videos. For both attention conditions, the scene environment and movements exacerbates or not UAV biases. For instance, in sequences *car 2-9* (fourth row), the aircraft is following cars on a road. Associated average human saliency maps display the shape of the road and its direction, i.e. vertical route for all and roundabout for *car7*. *Car 14* (third row), a semantically similar content except that it displays only one object on the road with a constant reframing (camera movement) which keeps the car at the same location, presents an average human saliency map centered on the tracked object.

Original database-specific center bias. We can observe that a center bias is present in VIRAT sequences, while videos from other datasets, namely UAV123 and DTB70, do not present this systematic bias. The original resolution of content and the experimental setup are possibly the sources of this

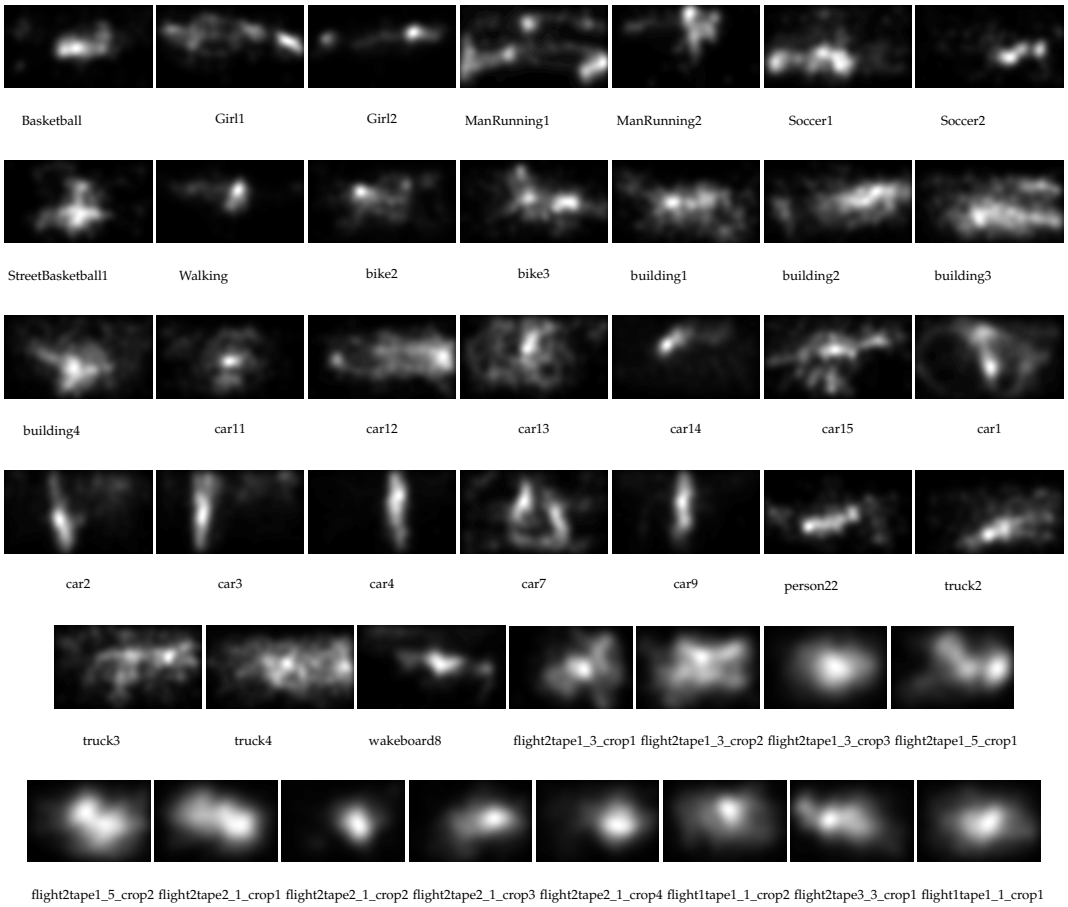


Figure 7. Average saliency maps for all sequences of *EyeTrackUAV2* dataset, generated from D scenario, for the free-viewing condition.

result. Indeed, the proportion of content seen at once is not the same for all sequences: 1,19% of a VIRAT content is seen per degree of visual angle, whereas it is 0,44% for the two other original databases. VIRAT saliency maps are thus smoother, which results in higher chances to present a center bias. To verify this assumption based on qualitative assessment, we have computed the overall human saliency maps for sequences coming from original dataset, namely DTB70, UAV123 and VIRAT. These maps are shown in Figure 9. VIRAT saliency maps are much more concentrated and centered. This corroborates that biases can be original-database-specific.

Task-related human saliency maps are more spread out. Task-based saliency maps cover more content when compared to free-viewing condition for most sequences (e.g. in about 58% of videos such as *Basketball, car11, car2, and wakeboard*). This behavior is also illustrated in Figure 9. We correlate this response with the object detection task. Visual search scanning implies an extensive exploration of the content. However, 21% of the remaining sequences (i.e. *soccer1, bike2-3, building 1-2, car1,15, and truck3-4*) show less discrepancies in the task-viewing condition than in free-viewing condition. We do not find correlation between such behavior and sequences characteristics given in Table 1. This leaves room for further exploration of differences between task-based and free viewing attention.

Overall, there is no generalization of center bias for UAV content. As stated earlier, we do not observe a systematic center bias, except for VIRAT sequences. This is especially true for task-related viewing. However, we observe specific patterns. Indeed, vertical and horizontal potatoe-shaped saliency areas are quite present in average human saliency maps of *EyeTrackUAV2*. Such patterns are also visible in UAV2 and DTB70 overall mean maps, especially in task-viewing condition. This indicates future axes of developments for UAV saliency-based applications. For instance, instead of

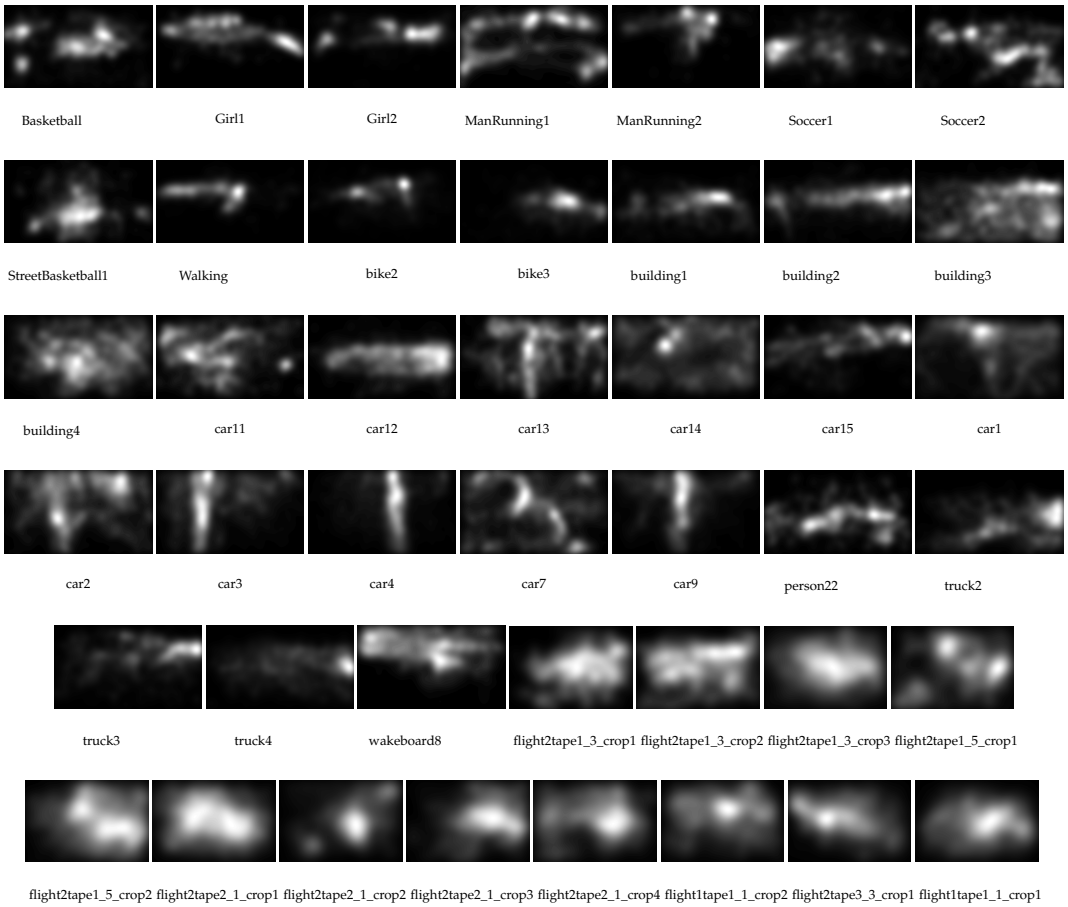


Figure 8. Average saliency maps for all sequences of *EyeTrackUAV2* dataset, generated from D scenario, for the task-viewing condition.

using a center bias, one may introduce priors as a set of prevalent saliency area shapes with different directions and sizes [81].

4.2.2. Quantitative evaluation of the central bias in UAV videos

To go further into content-dependencies, we investigate quantitatively the similarity of dominant-eye-generated saliency maps, called ground truth in the remaining, with a pre-defined center bias. Figure 10 presents the center bias baseline created in this purpose as suggested in [25,75] .

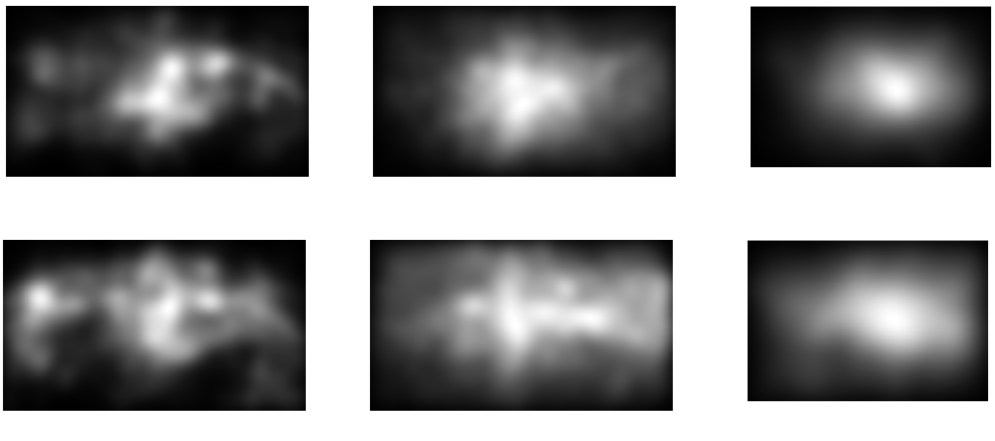
We performed the evaluation based on four well-used quality metrics: CC, SIM, KL, and IG. Results are presented in Table 7. They support the observations we made in the previous section. Overall scores do not reveal a high similarity with the center prior (e.g. maximum CC and SIM of about 0.5, high KL and IG). On the other hand, we observe content-specific center prior in UAV123 and DTB70. For instance, videos more prone to center bias includes sequences extracted from VIRAT and *building1,3,4*, and *car13*. On the contrary, sequences *Girl1-2, ManRunning1-2, Walking, car4*, and *wakeboard8* are not likely to present center bias. This confirms there is no generalization of center bias for UAV content. Regarding differences between free-viewing and task-viewing conditions, results are inconclusive as no systematic behavior is clearly visible from this analysis.

5. Conclusion

UAV imaging modifies the perceptual clues of typical scenes due to its bird point of view, the presence of camera movements and the high distance and angle to the scene. For instance, low-level visual features and size of objects changes and depth information is flattened or disappears (e.g.

	FV				Task			
	CC ↑	SIM ↑	KL ↓	IG ↓	CC ↑	SIM ↑	KL ↓	IG ↓
VIRAT_09152008flight2tape1_3_crop1	**0,50**	**0,48**	**7,17**	**1,53**	**0,46**	**0,48**	**6,85**	**1,62**
VIRAT_09152008flight2tape1_3_crop2	**0,49**	**0,52**	**5,59**	**1,50**	**0,36**	**0,48**	**6,42**	**1,75**
VIRAT_09152008flight2tape1_3_crop3	**0,46**	**0,43**	**8,46**	**1,91**	**0,37**	**0,43**	**7,98**	**1,99**
VIRAT_09152008flight2tape1_5_crop1	0,27	**0,38**	**9,77**	**2,29**	0,18	**0,36**	**10,14**	**2,49**
VIRAT_09152008flight2tape1_5_crop2	**0,42**	**0,44**	**8,05**	**1,90**	**0,30**	**0,45**	**7,41**	**1,87**
VIRAT_09152008flight2tape2_1_crop1	**0,41**	**0,39**	**9,34**	**2,05**	**0,38**	**0,42**	**8,55**	**1,97**
VIRAT_09152008flight2tape2_1_crop2	**0,40**	**0,35**	**10,90**	**2,50**	**0,32**	**0,42**	**8,01**	**2,01**
VIRAT_09152008flight2tape2_1_crop3	**0,42**	**0,40**	**9,46**	**2,11**	**0,28**	**0,39**	**9,30**	**2,24**
VIRAT_09152008flight2tape2_1_crop4	**0,36**	**0,36**	**10,35**	**2,34**	**0,28**	**0,38**	**9,79**	**2,30**
VIRAT_09152008flight2tape3_3_crop1	**0,42**	**0,43**	**8,16**	**1,96**	**0,35**	**0,43**	**7,84**	**2,03**
VIRAT_09162008flight1tape1_1_crop1	**0,47**	**0,45**	**7,76**	**1,80**	**0,37**	**0,42**	**8,40**	**2,00**
VIRAT_09162008flight1tape1_1_crop2	**0,40**	**0,40**	**9,14**	**2,14**	**0,27**	**0,40**	**8,91**	**2,22**
UAV123_bike2	**0,39**	**0,34**	**11,51**	**2,43**	**0,34**	0,29	13,21	2,82
UAV123_bike3	**0,39**	**0,34**	**11,71**	**2,37**	**0,29**	0,26	14,34	2,96
UAV123_building1	**0,40**	**0,37**	**10,64**	**2,18**	**0,32**	0,31	12,74	**2,69**
UAV123_building2	0,30	**0,33**	**11,89**	**2,43**	0,18	0,27	13,87	3,06
UAV123_building3	0,27	**0,34**	**11,50**	**2,42**	0,17	**0,32**	**11,82**	**2,56**
UAV123_building4	**0,39**	**0,36**	**10,82**	**2,20**	**0,35**	**0,39**	**9,72**	**2,10**
UAV123_car11	**0,37**	**0,32**	12,37	**2,58**	0,21	0,30	12,68	**2,67**
UAV123_car12	0,21	0,28	13,35	2,80	**0,26**	0,29	13,12	2,69
UAV123_car13	0,30	**0,34**	**11,48**	**2,39**	0,20	**0,33**	**11,50**	**2,44**
UAV123_car14	0,20	0,25	14,47	3,16	0,12	0,31	12,28	2,71
UAV123_car15	**0,31**	**0,34**	**11,52**	**2,47**	0,10	0,30	12,70	2,81
UAV123_car1	0,21	0,26	14,33	3,10	0,13	0,30	12,61	2,77
UAV123_car2	0,22	0,27	13,91	3,02	0,13	0,30	12,68	2,80
UAV123_car3	0,16	0,24	14,77	3,19	0,14	0,28	13,39	2,93
UAV123_car4	0,22	0,20	16,27	3,55	0,20	0,24	14,76	3,23
UAV123_car7	0,22	0,23	15,11	3,16	0,11	0,28	13,13	2,92
UAV123_car9	0,26	0,23	15,41	3,27	0,21	0,28	13,69	2,86
UAV123_person22	**0,35**	0,31	12,44	**2,60**	**0,27**	0,31	12,45	2,68
UAV123_truck2	0,27	**0,32**	12,29	**2,56**	0,09	0,27	13,66	3,01
UAV123_truck3	0,27	**0,35**	**11,14**	**2,34**	0,12	0,31	12,23	2,73
UAV123_truck4	0,29	**0,36**	**10,71**	**2,34**	0,16	0,29	13,18	3,03
UAV123_wakeboard8	0,23	0,21	15,91	3,45	0,11	0,24	14,93	3,29
DTB70_Basketball	**0,38**	0,27	14,13	2,89	**0,30**	0,31	12,30	**2,59**
DTB70_Girl1	0,16	0,28	13,47	2,90	0,15	0,25	14,54	3,18
DTB70_Girl2	0,20	0,20	16,04	3,60	0,19	0,23	15,04	3,34
DTB70_ManRunning1	0,02	0,16	17,45	4,09	0,00	0,20	16,11	3,73
DTB70_ManRunning2	0,12	0,13	18,40	4,31	0,10	0,15	17,99	4,24
DTB70_Soccer1	0,21	0,26	14,23	3,04	0,17	0,26	14,03	3,18
DTB70_Soccer2	0,21	0,22	15,56	3,33	0,22	**0,32**	**11,86**	**2,69**
DTB70_StreetBasketball1	**0,33**	0,26	14,29	2,94	**0,28**	0,26	14,29	3,00
DTB70_Walking	0,29	0,20	16,14	3,51	**0,27**	0,22	15,81	3,51
mean	0,31	0,32	12,27	2,67	0,23	0,32	12,01	2,69

Table 7. Comparison of saliency maps with the center bias presented in Figure 10. Are displayed in red the numbers over (or under for KL and IG) measures average, indicated in the last row.



(a) (b) (c)

Figure 9. Overall average saliency maps per original dataset, generated from D scenario, in free-viewing (top-row) and Task-viewing (bottom row) for original datasets: (a) DTB70; (b) UAV123; (c) VIRAT.

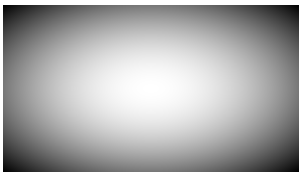


Figure 10. Center prior baseline.

presence of sky). To understand observers' behaviors toward these new features, especially in terms of visual attention and deployment, there is a need for large-scale eye-tracking databases for saliency in UAV videos. This dataset is also a key factor in the field of computational models of visual attention, in which large scale datasets are required to train the latest generation of deep-based models.

This need is even stronger with the fast expansion of applications related to UAVs, for leisure and professional civilian activities and a wide range of military services. Combining UAV imagery with one of the most dynamic research fields in vision, namely salience, is highly promising, especially for videos that are gaining more and more attention these last years.

This work addresses the need for such a dedicated dataset. An experimental process has been designed in order to build a new dataset, *EyeTrackUAV2*. Gaze data were collected during the observation of UAV videos under controlled laboratory conditions for both free viewing and object-detection surveillance task conditions. Gaze positions have been collected on 30 participants for each attention condition, on 43 UAV videos in 30 fps, 1280x720 or 720x480, consisting in 42 241 frames and 1408 seconds. Overall, 1 239 157 fixations in free-viewing and 1 269 433 in task-viewing were extracted from the dominant eye positions. Test stimuli were carefully selected from three original datasets, i.e. UAV123, VIRAT, and DTB70, to be representative as much as possible of the UAV ecosystem. Accordingly, they present variations in terms of environments, camera movement, size of objects, aircraft flight heights and angles to the ground, video duration, resolution, and quality. Also, three sequences were recorded in infra-red.

The collected gaze data were analyzed and transformed into fixation and saccade eye movements using an I-DT based identification algorithm. Moreover, the eye-tracking system high frequency of acquisition enabled the production of saliency maps for each experimental frame of the examined video stimuli directly from raw data. The dataset is publicly available and includes, for instance, raw binocular eye positions, fixation, and saliency maps generated from the dominant eye and both eyes information.

Then, we further characterized the dataset considering two different aspects. On one hand, six scenarios, namely binocular, both eyes, dominant eye, non-dominant eye, left, and right can be envisioned to generate human saliency maps. We wondered whether a scenario should be favored over another or not. Comparisons of scenarios have been conducted, first in terms of the mean of absolute errors for eye positions, and secondly on six typical saliency metrics for saliency maps. Results indicate that the cyclopean theory prevails over the ocular dominance theory in saliency. That means that information of both eyes should be favored to study saliency. If not possible, choosing information from the dominant eye allows us to commit fewer errors when compared to other one-eye scenarios. On the other hand, we notice that conventional biases in saliency do not necessarily apply to UAV content. Indeed, the center bias is not systematic in UAV sequences. This bias is content-dependent as well as and task-condition-dependent. We observed new prior patterns that must be examined in the future.

In conclusion, the *EyeTrackUAV2* dataset enables in-depth studies of visual attention through the exploration of new salience biases and prior patterns. It establishes in addition a solid basis on which dynamic salience for UAV imaging can build upon, in particular for the development of deep-learning saliency models.

6. Acknowledgment

The presented work is funded by the ongoing research project ANR ASTRID DISSOCIE (Automated Detection of SaliencieS from Operators' Point of View and Intelligent Compression of DronE videos) referenced as ANR-17-ASTR-0009. Specifically, the LS2N team ran the experiment, created and made available the *EyeTrackUAV2* dataset. The Univ Rennes team added binocular and both-eyed scenarios information to the dataset, conducted analyses, and reported it.

1. Zhao, Y.; Ma, J.; Li, X.; Zhang, J. Saliency detection and deep learning-based wildfire identification in UAV imagery. *Sensors* **2018**, *18*, 712.
2. van Gemert, J.C.; Verschoor, C.R.; Mettes, P.; Epema, K.; Koh, L.P.; Wich, S. Nature conservation drones for automatic localization and counting of animals. Workshop at the European Conference on Computer Vision. Springer, 2014, pp. 255–270.
3. Postema, S. News Drones: An Auxiliary Perspective, 2015.
4. Agbeyangi, A.O.; Odiete, J.O.; Olorunlomerue, A.B. Review on UAVs used for aerial surveillance. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)* **2016**, *3*, 5713–5719.
5. Lee-Morrison, L. *State of the Art Report on Drone-Based Warfare*; Citeseer, 2014.
6. Zhou, Y.; Tang, D.; Zhou, H.; Xiang, X.; Hu, T. Vision-Based Online Localization and Trajectory Smoothing for Fixed-Wing UAV Tracking a Moving Target. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.
7. Zhu, P.; Du, D.; Wen, L.; Bian, X.; Ling, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; others. VisDrone-VID2019: The Vision Meets Drone Object Detection in Video Challenge Results. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.
8. Aguilar, W.G.; Luna, M.A.; Moya, J.F.; Abad, V.; Ruiz, H.; Parra, H.; Angulo, C. Pedestrian detection for UAVs using cascade classifiers and saliency maps. International Work-Conference on Artificial Neural Networks. Springer, 2017, pp. 563–574.
9. Dang, T.; Khattak, S.; Papachristos, C.; Alexis, K. Anomaly Detection and Cognizant Path Planning for Surveillance Operations using Aerial Robots. 2019 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE, 2019, pp. 667–673.
10. Edney-Browne, A. Vision, visuality, and agency in the US drone program. *Technology and Agency in International Relations* **2019**, p. 88.
11. Krassanakis, V.; Perreira Da Silva, M.; Ricordel, V. Monitoring Human Visual Behavior during the Observation of Unmanned Aerial Vehicles (UAVs) Videos. *Drones* **2018**, *2*, 36.

12. Papachristos, C.; Khattak, S.; Mascarich, F.; Dang, T.; Alexis, K. Autonomous Aerial Robotic Exploration of Subterranean Environments relying on Morphology–aware Path Planning. 2019 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE, 2019, pp. 299–305.
13. Itti, L.; Koch, C. Computational modelling of visual attention. *Nature reviews neuroscience* **2001**, *2*, 194.
14. Katsuki, F.; Constantinidis, C. Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist* **2014**, *20*, 509–521.
15. Krasovskaya, S.; MacInnes, W.J. Salience Models: A Computational Cognitive Neuroscience Review. *Vision* **2019**, *3*, 56.
16. Kummerer, M.; Wallis, T.S.; Bethge, M. Saliency benchmarking made easy: Separating models, maps and metrics. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 770–787.
17. Riche, N.; Duvinage, M.; Mancas, M.; Gosselin, B.; Dutoit, T. Saliency and human fixations: State-of-the-art and study of comparison metrics. Proceedings of the IEEE international conference on computer vision, 2013, pp. 1153–1160.
18. Guo, C.; Zhang, L. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE transactions on image processing* **2009**, *19*, 185–198.
19. Jain, S.D.; Xiong, B.; Grauman, K. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, 2017, pp. 2117–2126.
20. Wang, W.; Shen, J.; Shao, L. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing* **2017**, *27*, 38–49.
21. Li, G.; Xie, Y.; Wei, T.; Wang, K.; Lin, L. Flow guided recurrent neural encoder for video salient object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3243–3252.
22. Le Meur, O.; Coutrot, A.; Liu, Z.; Rämä, P.; Le Roch, A.; Helo, A. Visual attention saccadic models learn to emulate gaze patterns from childhood to adulthood. *IEEE Transactions on Image Processing* **2017**, *26*, 4777–4789.
23. Perrin, A.F.; Zhang, L.; Le Meur, O. How well current saliency prediction models perform on UAVs videos? International Conference on Computer Analysis of Images and Patterns. Springer, 2019, pp. 311–323.
24. Paglin, M.; Rufolo, A.M. Heterogeneous human capital, occupational choice, and male-female earnings differences. *Journal of Labor Economics* **1990**, *8*, 123–144.
25. Le Meur, O.; Le Callet, P.; Barba, D.; Thoreau, D. A coherent computational approach to model bottom-up visual attention. *IEEE transactions on pattern analysis and machine intelligence* **2006**, *28*, 802–817.
26. Ehinger, K.A.; Hidalgo-Sotelo, B.; Torralba, A.; Oliva, A. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition* **2009**, *17*, 945–978.
27. Liu, H.; Heynderickx, I. Studying the added value of visual attention in objective image quality metrics based on eye movement data. 2009 16th IEEE international conference on image processing (ICIP). IEEE, 2009, pp. 3097–3100.
28. Judd, T.; Durand, F.; Torralba, A. A benchmark of computational models of saliency to predict human fixations, 2012.
29. Ma, K.T.; Sim, T.; Kankanhalli, M. VIP: A unifying framework for computational eye-gaze research. International Workshop on Human Behavior Understanding. Springer, 2013, pp. 209–222.
30. Koehler, K.; Guo, F.; Zhang, S.; Eckstein, M.P. What do saliency models predict? *Journal of vision* **2014**, *14*, 14–14.
31. Borji, A.; Itti, L. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581* **2015**.
32. Bylinskii, Z.; Isola, P.; Bainbridge, C.; Torralba, A.; Oliva, A. Intrinsic and extrinsic effects on image memorability. *Vision research* **2015**, *116*, 165–178.
33. Fan, S.; Shen, Z.; Jiang, M.; Koenig, B.L.; Xu, J.; Kankanhalli, M.S.; Zhao, Q. Emotional attention: A study of image sentiment and visual attention. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7521–7531.
34. McCamy, M.B.; Otero-Millan, J.; Di Stasi, L.L.; Macknik, S.L.; Martinez-Conde, S. Highly informative natural scene regions increase microsaccade production during visual scanning. *Journal of neuroscience* **2014**, *34*, 2956–2966.

35. Gitman, Y.; Erofeev, M.; Vatolin, D.; Andrey, B.; Alexey, F. Semiautomatic visual-attention modeling and its application to video compression. 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014, pp. 1105–1109.
36. Coutrot, A.; Guyader, N. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of vision* **2014**, *14*, 5–5.
37. Coutrot, A.; Guyader, N. An efficient audiovisual saliency model to predict eye positions when looking at conversations. 2015 23rd European Signal Processing Conference (EUSIPCO). IEEE, 2015, pp. 1531–1535.
38. Wang, W.; Shen, J.; Xie, J.; Cheng, M.M.; Ling, H.; Borji, A. Revisiting video saliency prediction in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence* **2019**.
39. Oh, S.; Hoogs, A.; Perera, A.; Cuntoor, N.; Chen, C.C.; Lee, J.T.; Mukherjee, S.; Aggarwal, J.; Lee, H.; Davis, L.; others. A large-scale benchmark dataset for event recognition in surveillance video. CVPR 2011. IEEE, 2011, pp. 3153–3160.
40. Layne, R.; Hospedales, T.M.; Gong, S. Investigating open-world person re-identification using a drone. European Conference on Computer Vision. Springer, 2014, pp. 225–240.
41. Bonetto, M.; Korshunov, P.; Ramponi, G.; Ebrahimi, T. Privacy in mini-drone based video surveillance. 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). IEEE, 2015, Vol. 4, pp. 1–6.
42. Shu, T.; Xie, D.; Rothrock, B.; Todorovic, S.; Chun Zhu, S. Joint inference of groups, events and human roles in aerial videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4576–4584.
43. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. European conference on computer vision. Springer, 2016, pp. 445–461.
44. Robicquet, A.; Sadeghian, A.; Alahi, A.; Savarese, S. Learning social etiquette: Human trajectory understanding in crowded scenes. European conference on computer vision. Springer, 2016, pp. 549–565.
45. Li, S.; Yeung, D.Y. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. Thirty-First AAAI Conference on Artificial Intelligence, 2017.
46. Barekatain, M.; Martí, M.; Shih, H.F.; Murray, S.; Nakayama, K.; Matsuo, Y.; Prendinger, H. Okutama-action: An aerial view video dataset for concurrent human action detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 28–35.
47. Hsieh, M.R.; Lin, Y.L.; Hsu, W.H. Drone-based object counting by spatially regularized regional proposal network. Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4145–4153.
48. Ribeiro, R.; Cruz, G.; Matos, J.; Bernardino, A. A dataset for airborne maritime surveillance environments. *IEEE Trans. Circuits Syst. Video Technol* **2017**.
49. Hsu, H.J.; Chen, K.T. DroneFace: an open dataset for drone research. Proceedings of the 8th ACM on Multimedia Systems Conference. ACM, 2017, pp. 187–192.
50. Božić-Štulić, D.; Marušić, Ž.; Gotovac, S. Deep Learning Approach in Aerial Imagery for Supporting Land Search and Rescue Missions. *International Journal of Computer Vision* **2019**, pp. 1–23.
51. Fu, K.; Li, J.; Shen, H.; Tian, Y. How drones look: Crowdsourced knowledge transfer for aerial video saliency prediction. *arXiv preprint arXiv:1811.05625* **2018**.
52. Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437* **2018**.
53. Nyström, M.; Andersson, R.; Holmqvist, K.; Van De Weijer, J. The influence of calibration method and eye physiology on eyetracking data quality. *Behavior research methods* **2013**, *45*, 272–288.
54. ITU-T RECOMMENDATION, P. Subjective video quality assessment methods for multimedia applications. *International telecommunication union* **2008**.
55. Rec, I. BT. 710-4. *Subjective assessment methods for image quality in high-definition television* **1998**.
56. Cornelissen, F.W.; Peters, E.M.; Palmer, J. The Eyelink Toolbox: eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers* **2002**, *34*, 613–617.
57. Rec, I. BT. 500-13. *Methodology for the subjective assessment of the quality of television pictures* **2012**.
58. Wandell, B.; Thomas, S. Foundations of vision. *Psyccritiques* **1997**, *42*.
59. Le Meur, O.; Liu, Z. Saccadic model of eye movements for free-viewing condition. *Vision research* **2015**, *116*, 152–164.

60. Le Meur, O.; Baccino, T. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods* **2013**, *45*, 251–266.
61. Guznov, S.; Matthews, G.; Warm, J.S.; Pfahler, M. Training techniques for visual search in complex task environments. *Human factors* **2017**, *59*, 1139–1152.
62. Shah, M.; Javed, O.; Shafique, K. Automated visual surveillance in realistic scenarios. *IEEE MultiMedia* **2007**, *14*, 30–39.
63. Snellen, H. *Test-types for the determination of the acuteness of vision*; Williams and Norgate, 1868.
64. Ishihara, S. *Test for colour-blindness*; Kanehara Tokyo, Japan, 1987.
65. Salvucci, D.D.; Goldberg, J.H. Identifying fixations and saccades in eye-tracking protocols. Proceedings of the 2000 symposium on Eye tracking research & applications. ACM, 2000, pp. 71–78.
66. Krassanakis, V.; Filippakopoulou, V.; Nakos, B. EyeMMV toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification. *Journal of eye movement research* **2014**.
67. Krassanakis, V.; Misthos, L.M.; Menegaki, M. LandRate toolbox: An adaptable tool for eye movement analysis and landscape rating. Eye Tracking for Spatial Research, Proceedings of the 3rd International Workshop. ETH Zurich, 2018.
68. Krassanakis, V.; Filippakopoulou, V.; Nakos, B. Detection of moving point symbols on cartographic backgrounds. *Journal of Eye Movement Research* **2016**, *9*.
69. Ooms, K.; Krassanakis, V. Measuring the Spatial Noise of a Low-Cost Eye Tracker to Enhance Fixation Detection. *Journal of Imaging* **2018**, *4*, 96.
70. Bindemann, M. Scene and screen center bias early eye movements in scene viewing. *Vision research* **2010**, *50*, 2577–2587.
71. Cui, Y.; Hondzinski, J.M. Gaze tracking accuracy in humans: Two eyes are better than one. *Neuroscience letters* **2006**, *396*, 257–262.
72. Holmqvist, K.; Nyström, M.; Mulvey, F. Eye tracker data quality: what it is and how to measure it. Proceedings of the symposium on eye tracking research and applications. ACM, 2012, pp. 45–52.
73. Hooge, I.T.; Holleman, G.A.; Haukes, N.C.; Hessels, R.S. Gaze tracking accuracy in humans: One eye is sometimes better than two. *Behavior Research Methods* **2018**, pp. 1–10.
74. Bylinskii, Z.; Judd, T.; Borji, A.; Itti, L.; Durand, F.; Oliva, A.; Torralba, A. Mit saliency benchmark, 2015.
75. Bylinskii, Z.; Judd, T.; Oliva, A.; Torralba, A.; Durand, F. What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2019**, *41*, 740–757.
76. Tseng, P.H.; Carmi, R.; Cameron, I.G.; Munoz, D.P.; Itti, L. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision* **2009**, *9*, 4–4.
77. Van Opstal, A.; Hepp, K.; Suzuki, Y.; Henn, V. Influence of eye position on activity in monkey superior colliculus. *Journal of Neurophysiology* **1995**, *74*, 1593–1610.
78. Tatler, B.W. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision* **2007**, *7*, 4–4.
79. Howard, I.P.; Rogers, B. Depth perception. *Stevens Handbook of Experimental Psychology* **2002**, *6*, 77–120.
80. Foulsham, T.; Kingstone, A.; Underwood, G. Turning the world around: Patterns in saccade direction vary with picture orientation. *Vision research* **2008**, *48*, 1777–1790.
81. Le Meur, O.; Coutrot, A. Introducing context-dependent and spatially-variant viewing biases in saccadic models. *Vision research* **2016**, *121*, 72–84.