DR(eye)VE: a Dataset for Attention-Based Tasks with Applications to Autonomous and Assisted Driving

Stefano Alletto*, Andrea Palazzi*, Francesco Solera*, Simone Calderara and Rita Cucchiara
University of Modena and Reggio Emilia
http://imagelab.ing.unimore.it/dreyeve

Abstract

Autonomous and assisted driving are undoubtedly hot topics in computer vision. However, the driving task is extremely complex and a deep understanding of drivers' behavior is still lacking. Several researchers are now investigating the attention mechanism in order to define computational models for detecting salient and interesting objects in the scene. Nevertheless, most of these models only refer to bottom up visual saliency and are focused on still images. Instead, during the driving experience the temporal nature and peculiarity of the task influence the attention mechanisms, leading to the conclusion that real life driving data is mandatory. In this paper we propose a novel and publicly available dataset acquired during actual driving. Our dataset, composed by more than 500,000 frames, contains drivers' gaze fixations and their temporal integration providing task-specific saliency maps. Geo-referenced locations, driving speed and course complete the set of released data. To the best of our knowledge, this is the first publicly available dataset of this kind and can foster new discussions on better understanding, exploiting and reproducing the driver's attention process in the autonomous and assisted cars of future generations.

1. Introduction

Autonomous and assisted driving have recently gained increasing momentum in the computer vision community. With the advent of deep learning, many tasks involving visual understanding –something which cannot be eluded in driving– have reached human-level performance, and sometimes overtaken it. Examples are pedestrian and vehicle detection and tracking, lanes or road signs recognition and, ultimately, semantic segmentation, where each pixel gets a label according to what it represents [5, 37]. All of these achievements are great examples of *subtasks* to autonomous and assisted driving, but we must not forget that the utmost goal is (better) driving itself. Do humans really need to detect all pedestrians or recognize all signs to drive? Do humans really need to label the whole scene?

Figure 1. An exemplar frame from our dataset. From left to right, from up to bottom: car-mounted view, driver's point of view, gaze map overlay and geo-referenced course.

In widely accepted psychological studies on the topic, the connection between driving, attention and gaze has been explored [27], negatively answering the above questions. It is known that humans' selective attention is a constraint required by the limited amount of resources available to our brain. Hence, it is still debatable if this approach may also bring benefits to visual computing models where the computational resources can be raised by adopting advanced performant hardware (e.g. GPUs, clusters). Nevertheless, the act of driving combines attention mechanisms influenced by the driver past experience, the temporal nature of the task and strong contextual constraints. As a result, we can drive much more safely and effectively than any automated system. One of the most relevant open questions in the field is to establish whether autonomous cars could benefit from attention-like mechanisms as well. Unluckily, this topic is under-investigated in computer vision and the lack of a realistic experimental framework does not help.

Our main contribution is a new dataset available to the community, depicted in Fig. 1. We recorded more than six hours and 500,000 frames of driving sequences in different

* equal contribution

traffic and weather conditions. For every frame, we also acquire the driver gaze through an accurate eye tracking device. Additionally, to favor the car point of view, we project gaze information on a HD quality video recorded from a roof-mounted camera. Given the subjective nature of both attention and driving, experimental design has played a crucial role in preparing the dataset and rule out spurious correlation between driver, weather, traffic, daytime and scenario.

At a computational level, human attention and eye fixation are typically modeled through the concept of *visual saliency*. Most of the literature on visual saliency focuses on filtering, selecting and synthesizing task dependent features for automatic object recognition. Nevertheless, the majority of experiments are constructed in controlled environments (*e.g.* laboratory settings) and on sequences of unrelated images [30, 4, 15]. Conversely, our dataset has been collected "on the road" and it exhibits the following features:

- *It is public and open.* It provides hours of driving videos that can be used for understanding the attention phenomena;
- *It is task and context dependent.* According to the psychological studies on attention, data are collected during a real driving experience thus being as much realistic as possible;
- *It is precise and scientifically solid.* We use high end attention recognition instruments, in conjunction with camera data and GPS information.

We believe that our proposal can be useful in several contexts aimed at understanding the driving phenomenon. It can be applied to identify and collect new features tailored for the driving experience (by analogy with what recently studied for video action recognition [21]). It can help understanding the influence of motion and semantics in salient object detection [26, 32]. It can foster the creation of new driver-centric visual onthologies, and as well serve the purpose to better understand how driver past experience affects the importance of objects in the scene.

The paper is organized as follows. In Sec. 2, related works about computer vision and saliency are provided to frame the work in the current state of the art scenario. Sec. 3 describe the acquisition apparatus and protocol while Sec. 4 highlights the dataset features and peculiarities. Eventually, the paper is concluded with a discussion on the possible uses of the collected data.

2. Saliency and Gaze in Driving up to Now

Visual saliency determines how much each pixel of a scene attracts the observer's attention. This task can be approached by considering either a bottom up or a top down strategy. The former refers to data-driven saliency, as when a salient event pops out in the image. Here, visual discontinuity prevails and computational models focus on spotting these discontinuities by either clustering features or spotting the rarity of image regions either locally [24, 19] or globally [1, 34, 6]. On the opposite, top down saliency is task-driven, and refers to the objects characteristics which are relevant with respect to the ongoing task. At a glance, top-down computer vision models tend to exploit the semantic context in the saliency extraction process [29]. This is achieved by either fusing saliency maps at an increasing level of scale and abstraction [11], or injecting an a-priori model of the relevant object using tailored features or pretrained detectors [33, 10, 7].
Besides the aforementioned dichotomy between top down and bottom up saliency methods, deep networks have been employing the two approaches jointly to solve the task, achieving competitive results on public benchmarks [16, 17, 14].

In a broader sense, the literature on the topic agrees that video saliency falls in the latter category. Detecting saliency in videos is indeed a more difficult task because motion affects the attention mechanism driving the human gaze. Motion maps are usually fused with bottom up saliency maps by means of metric learning algorithms or supervised classifiers [38, 35]. In video saliency, motion has been computed by means of optical flow [38] or feature tracking [35].

2.1. Existing Datasets

Many image saliency datasets have been released in the past 5 years. They have driven most of the advancements in understanding the visual attention model and the computational mechanisms behind it. Most of the publicly available datasets are focusing on individual image saliency by capturing several users' fixation and integrating their spatial location by means of Gaussian filtering. Datasets can be distinguished between the ones annotated using an eye tracking system, such as the MIT saliency benchmark [4], and the ones where users click on images, like the SALICON dataset [15]. For a comprehensive list of datasets, the reader can refer to this recent survey [2].

While image datasets are publicly available and well established in the community, video saliency datasets are still lacking. Among the most important contributions to the field, we consider worth mentioning the *Action in the Eye* dataset [21], that consists in the largest video dataset providing human gaze and fixations during the task of action recognition. On the other hand, few driving datasets have been adopted for studying the attention phenomenon, with experiments conducted in laboratory settings and not made available to the community. In [25, 30], the fixations and saliency maps are acquired by simulating the driving experience. This setting is very limited and the observer atten-

tion can be driven by external factors independent from the driving task (*e.g.* monitor distance, normal attitude towards screen center and others) [27]. The few existing naturalistic in-car datasets [3, 23] are strictly designed asking the driver to accomplish a specific task (*e.g.* looking at people, traffic sign) and are not publicly available. Perhaps, the dataset proposed in [23] is the most similar to ours for naturalistic conditions and duration. Yet, it only records data for one driver, in two countryside scenarios and is not publicly available.

2.2. Saliency and Gaze for Assisted Driving

In the context of assisted driving, gaze and saliency inspection has been mainly studied in task specific environments and by acquiring the gaze using on screen images. In [30] object saliency is employed in order to avoid the *looked-but-failed-to-see* effect, by inspecting the attention of the driver towards pedestrian and motorbikes at T junctions. Bremond *et al.* [25] focus on enhancing the detection of traffic signs by exploiting visual saliency and a non-linear SVM classifier. This model was validated in a laboratory setting, pretending to drive a car and extended to a broader set of objects (*e.g.* pedestrian, bicycles) in a naturalistic experiment [3].

Recently, gaze has been studied in the context of preattentive driving, aiming at predicting the driver's next move by merely relying on its eye fixation [23]. When driver's gaze cannot be directly acquired through eye tracking systems, a set of independent works propose to inspect drivers' faces, using landmarks and predicting the head orientation [9, 28, 31]. While this mechanism has practical applications, there are no guarantees on the adherence of the results to the true gaze during the driving task.

In the current scenario, we believe that a publicly available dataset specifically tailored for the driving task and acquired during real driving conditions can significantly contribute to the research advancement in the field.

3. Apparatus and Acquisition protocol

To acquire information regarding the driver's gaze, we adopt the commercial *SMI ETG 2w* eye tracking glasses (ETG). Being head-mounted, they allow to fully capture the driver attention even under severe head pose changes, such as the ones that naturally occur during the drive. The device features a HD frontal camera acquiring at 720p/30fps, and two inner cameras solely devoted to tracking the user's pupil at 60fps. It provides information about the user's gaze in terms of eye fixations, saccade movements, blinks and pupil dilation. In order to ensure the highest possible gaze quality, 3-points calibration is performed before each recorded sequence to adapt to small changes in the ETG device position.

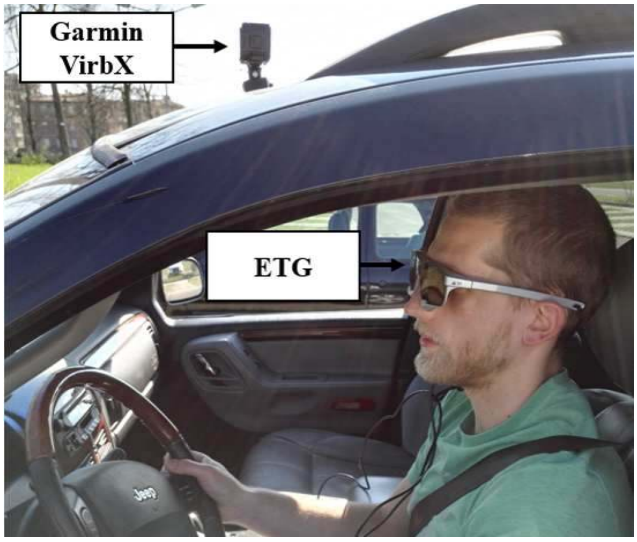

Figure 2. The acquisition rig featuring the head-mounted ETG and the car-mounted camera.

To collect videos from the car perspective, we adopt a roof-mounted *GARMIN VirbX* camera. It acquires videos at a resolution of 1080p/25fps, embeds an on-board GPS, accelerometer and gyroscope sensors and is waterproof, allowing to acquire video sequences under very different environmental conditions. Figure 2 illustrates the aforementioned acquisition rig.

During the acquisition phase, the two cameras are started simultaneously and the resulting videos are manually aligned to the frame in an offline stage to achieve the best possible synchronization. In order to re-project the gaze point on the video acquired by the car-mounted camera, local keypoints correspondences are exploited, resulting in the gaze information being present on both video sequences (see Section 4).

4. Dataset Description and Annotation

To overcome the lacks of the existing datasets described in Section 2, we acquire and publicly release 74 video sequences of 5 minutes each of actual driving experience, for a total of 555,000 frames[1]. Eight different drivers alternate during the recording process in order to smooth the bias given by each person's peculiar way of driving. To cover a wider range of scenarios, the videos are recorded in different areas of the city (downtown, countryside, highway) and present a broad reach of traffic conditions going from traffic-free to very cluttered situations. We also perform the recordings during completely diverse atmospheric conditions (sunny, cloudy and rainy) and in different times of the day, both at daytime and at night.

While the head-mounted ETG provides its own video se-

[1] http://imagelab.ing.unimore.it/dreyeve

Table 1. Table summarizing the different characteristics of the dataset.

# Videos	# Frames	Drivers	Weather conditions	Lighting	Gaze Info	Metadata	Camera POVs
74	555,000	8	sunny	day	raw fixations	GPS	driver (720p)
			cloudy	evening	gaze map	car speed	car (1080p)
			rainy	night	pupil dilation	car course	

quence, it is useful to project the acquired gaze position on the video acquired by the car-mounted camera. In fact, this camera features a significantly wider field of view (FoW) and can display fixations that are captured by the tracking device but not rendered by its video due to its limited FoW, such as the ones that occur when the driver peeks at something without moving his head. Since the two sequences have been manually aligned, they represent the same scene from two different but closely related perspectives and an homography transformation between the two can be employed to project the fixation points from one sequence to the other.

To estimate this transformation, Scale Invariant Feature Transform (SIFT) keypoints are extracted from the two frames [18] and a first, tentative nearest-neighbor matching is performed. Despite the sequence alignment, this matching step still produces outliers thus requiring robust techniques in the transformation estimation. For this purpose, the Random Sample Consensus (RANSAC) algorithm is employed [8]. Given four randomly selected corresponding points, the algorithm iteratively generates transformation hypotheses and selects the one that minimizes the reprojection error. Being H the selected homography matrix and P_g the fixation point on the ETG video (in homogeneous coordinates), its corresponding position on the car-mounted frame can be recovered as $P_c = H \times P_g$.

Psychological studies [20, 12] demonstrated that the scanpath of a scene is highly subjective and thus individual fixation points on a given frame may not be directly used to build a gaze map. Following this insight, we compute gaze maps at each frame that take into account all the fixations occurring in an interval of time centered on that frame. This allows to abstract from individual scanpaths and to obtain a gaze map that effectively covers the different parts of the scene that captured the driver's attention. To compute the map for a given frame F_m, a temporal window of $k = 25$ frames centered on F_m is selected. The choice of maps accumulating fixations over 25 frames follows [13] and ensures that the map accounts for the Inhibition of Return (IoR) mechanism [22], where broader maps would incur in redundant gaze information. By computing the homography transformation between each frame F_{m+i} with $i = -\frac{k}{2}, -\frac{k}{2}+1, \ldots, +\frac{k}{2}$ and F_m, the fixations occurring throughout the sequence are projected on the central frame. To obtain a smooth map, spatio-temporal Gaussian smoothing $G(\sigma_s, \sigma_t)$ is performed on the frame F_m, with $\sigma_s = 200$ pixels being the spatial variance, and $\sigma_t = \frac{k}{2}$ being the temporal variance. Figure 3 shows an example of gaze map obtained through this process.

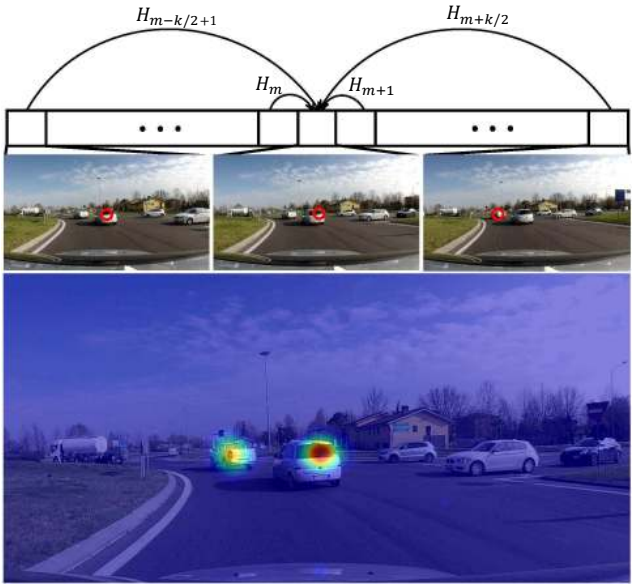

Figure 3. The resulting gaze map from a 25 frames sequence

5. Discussion and Open Questions

In this section we pose a few challenges and opportunities unlocked by the availability of the proposed dataset to the computer vision community. According to a qualitative analysis, it appears that when using an image based saliency prediction method (*e.g.* [36], which achieves state of the art performance on [4]), the regions of interest heavily rely on visual discontinuities resulting in fairly different attention maps with respect to the driver actual intentions, Figure 4 fourth and fifth columns. While this difference has not yet been quantitatively studied, it raises a set of open questions that we believe of interest for the computer vision community. Investigating the following topics (and possibly achieving positive answers) may consequently help pushing forward the field of assisted and autonomous driving.

Can driver's gaze be predicted?

Despite a large body of psychological literature, the computer vision community has not yet seen effective computational models able to predict human gaze while driving. In particular, the temporal nature of the driving task has never

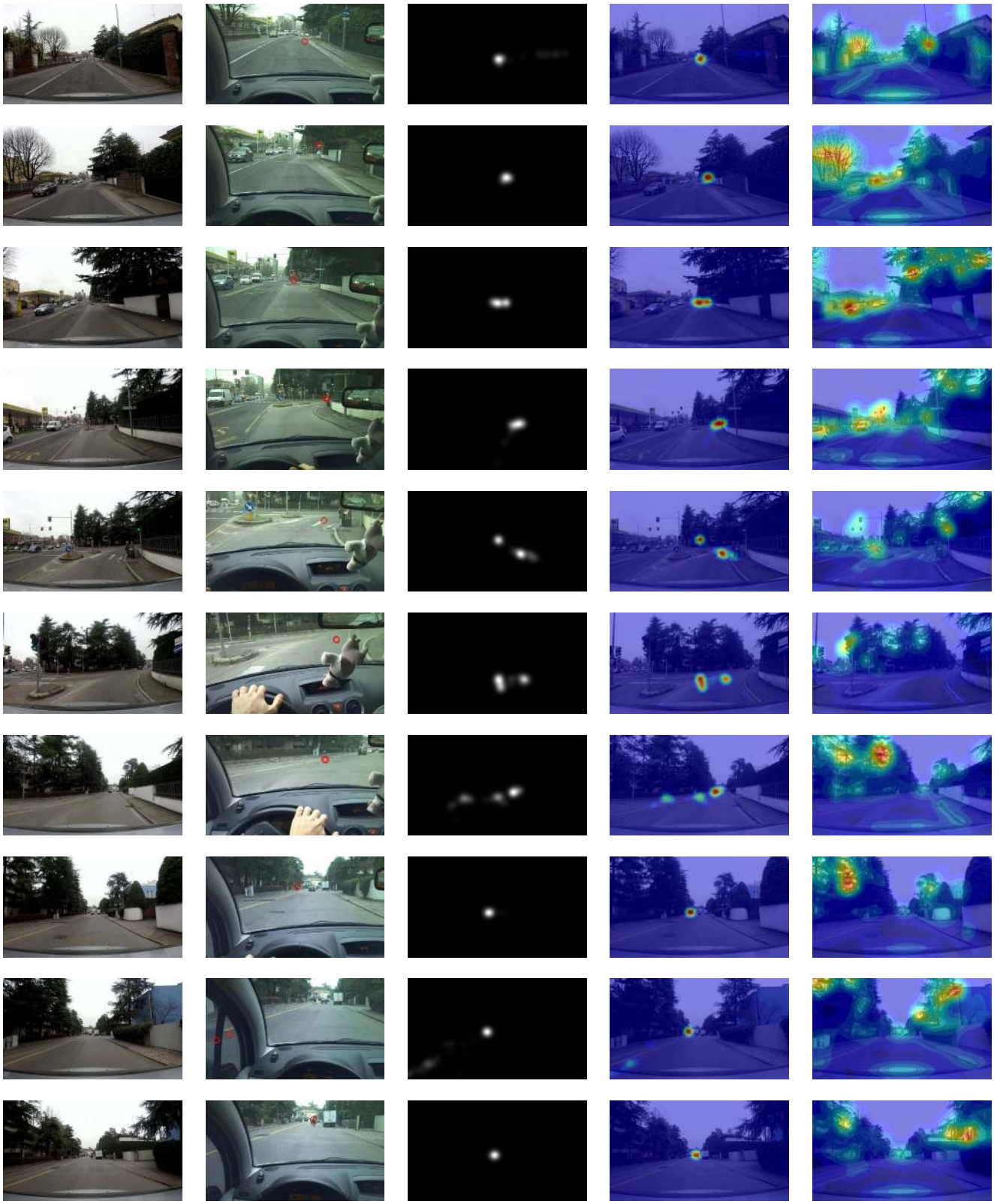

Figure 4. An example sequence taken from the dataset. Columns from left to right: Garmin VirbX frames, ETG frames with fixation information, the available gaze map, overlay between the frame and the gaze map, visual saliency predicted using [36]. From up to bottom: the temporal dimension of the video with 1 frame every 30 displayed.

been considered. In point of fact, we qualitatively observed that during red traffic lights and jams, visual saliency models trained on images could predict driver gaze quite accurately, [16, 4, 36]. Nevertheless, as driving speed increases, the amount of attention drivers dispose of at each instant decreases, resulting into very sparse and specific attention regions. Future models will need to take into account this behavior to provide reliable accuracy. Moreover, how easier is this task going to be if we were to feed the driver intentions (*e.g.* turn right in 10s) to the model?

Can driver's intentions be anticipated from gaze data?

Here we pose the opposite challenge to gaze prediction, that is whether we can build models that given video data and related gaze (true or predicted) are able to estimate the driver next move. These estimates can include the car turning angle, instantaneous speed, breaking events and so on. On top of this, the community may build systems able to exploit intentions prediction to alert the driver in dangerous situations.

Can gaze models be employed to enhance signalization and road safety?

While driving we only observe a small part of all the road signs, cars and traffic lights. In most of the cases, this is due to drivers' confidence about the path taken or irrelevant signalization with respect to driver current intentions. At the same time, overconfidence during driving may result in mistakes whenever signals change leading to possible dangerous situations. Local administrations can take advantage from gaze models to better decide how to place road signals and traffic lights. This is not a completely new line of work [25, 3], however the availability of a public dataset can serve as a unified benchmark for the research community.

Can gaze models help autonomous cars in planning better driving strategies?

Autonomous cars leverage on many different levels of structured information, ranging from lanes detection to semantic segmentation. Nevertheless, autonomous driving is ultimately a decision task. Can gaze information be yet another level of information to input to this decision process? Can human-like attention bring benefits to human-less vehicles? This is probably a far reaching question and we fully expect better experimental frameworks to be conceived in the future in order to answer it. Meanwhile, we make available the first dataset for the community to download and start tackling this challenge.

6. Conclusion

We propose a novel dataset that addresses the lack of public benchmarks concerning drivers' attention in real-world scenarios. It comes with pre-computed gaze maps and contextual information such as the car's speed and course. While focusing on the study of the driver's attention and gaze, we are planning to further extend its annotation to semantic labeling and driving actions. The dataset is freely available for academic research, along with the code used in the creation of gaze maps and annotation.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1597–1604, June 2009. 2

[2] A. Borji and L. Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581. 2

[3] R. Brémond, J.-M. Auberlet, V. Cavallo, L. Désiré, V. Faure, S. Lemonnier, R. Lobjois, and J.-P. Tarel. Where we look when we drive: A multidisciplinary approach. In *Proceedings of Transport Research Arena (TRA'14)*, Paris, France, 2014. http://perso.lcpc.fr/tarel.jean-philippe/publis/tra14b.html. 3, 6

[4] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. http://saliency.mit.edu/. 2, 4, 6

[5] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2015. 1

[6] M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, March 2015. 2

[7] L. Elazary and L. Itti. A bayesian model for efficient visual search and recognition. *Vision Research*, 50(14):1338 – 1352, 2010. Visual Search and Selective Attention. 2

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 4

[9] L. Fridman, P. Langhans, J. Lee, and B. Reimer. Driver gaze estimation without using eye movement. *CoRR*, abs/1507.04760, 2015. 3

[10] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant centersurround hypothesis for visual saliency. *Journal of Vision*, pages 1–18, 2008. 2

[11] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(10):1915–1926, Oct. 2012. 2

[12] R. Groner, F. Walder, and M. Groner. Looking at faces: Local and global aspects of scanpaths. *Advances in Psychology*, 22:523–533, 1984. 4

[13] J. M. Henderson. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504, 2003. 4

[14] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 262–270, Dec 2015. 2

[15] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[16] M. Kümmerer, L. Theis, and M. Bethge. Deep gaze I: boosting saliency prediction with feature maps trained on imagenet. *CoRR*, abs/1411.1045, 2014. 2, 6

[17] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 362–370, June 2015. 2

[18] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999. 4

[19] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, MULTIMEDIA '03, pages 374–381, New York, NY, USA, 2003. ACM. 2

[20] S. Mannan, K. Ruddock, and D. Wooding. Fixation sequences made during visual examination of briefly presented 2d images. *Spatial vision*, 11(2):157–178, 1997. 4

[21] S. Mathe and C. Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1408–1424, July 2015. 2

[22] M. I. Posner, R. D. Rafal, L. S. Choate, and J. Vaughan. Inhibition of return: Neural basis and function. *Cognitive neuropsychology*, 2(3):211–228, 1985. 4

[23] N. Pugeault and R. Bowden. How much of driving is preattentive? *IEEE Transactions on Vehicular Technology*, 64(12):5424–5438, Dec 2015. 3

[24] B. Schlkopf, J. Platt, and T. Hofmann. *Graph-Based Visual Saliency*, pages 545–552. MIT Press, 2007. 2

[25] L. Simon, J. P. Tarel, and R. Bremond. Alerting the drivers about road signs with poor visual saliency. In *Intelligent Vehicles Symposium, 2009 IEEE*, pages 48–53, June 2009. 3, 6

[26] N. Souly and M. Shah. Visual saliency detection using group lasso regularization in videos of natural scenes. *International Journal of Computer Vision*, 117(1):93–110, 2015. 2

[27] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5), May 2011. 1, 3

[28] A. Tawari and M. M. Trivedi. Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos. In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, pages 344–349, June 2014. 3

[29] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006. 2

[30] G. Underwood, K. Humphrey, and E. van Loon. Decisions about objects in real-world scenes are influenced by visual saliency before and during their inspection. *Vision Research*, 51(18):2031 – 2038, 2011. 2, 3

[31] F. Vicente, Z. Huang, X. Xiong, F. D. la Torre, W. Zhang, and D. Levi. Driver gaze tracking and eyes off the road detection system. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):2014–2027, Aug 2015. 3

[32] W. Wang, J. Shen, and F. Porikli. Saliency-aware geodesic video object segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3395–3402, June 2015. 2

[33] J. M. Wolfe, K. R. Cave, and S. L. Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception & Performance*, 1989. 2

[34] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th ACM International Conference on Multimedia*, MM '06, pages 815–824, New York, NY, USA, 2006. ACM. 2

[35] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th ACM International Conference on Multimedia*, MM '06, pages 815–824, New York, NY, USA, 2006. ACM. 2

[36] J. Zhang and S. Sclaroff. Exploiting surroundedness for saliency detection: A boolean map approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 2015. 4, 5, 6

[37] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)*, 2015. 1

[38] S.-h. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *AAAI*, 2013. 2