From Driver *Talk* To Future *Action*: Vehicle Maneuver Prediction by Learning from Driving Exam Dialogs

Alina Roitberg Simon Reiß Rainer Stiefelhagen

Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology
{firstname.lastname}@kit.edu

***Abstract*— A rapidly growing amount of content posted online inherently holds knowledge about concepts of interest, *i.e.* driver actions. We leverage methods at the intersection of vision and language to surpass costly annotation and present the first automated framework for anticipating driver intention by learning from recorded driving exam conversations. We query YouTube and collect a dataset of posted mock road tests comprising student-teacher dialogs and video data, which we use for learning to foresee the next maneuver without any additional supervision. However, instructional conversations give us very loose labels, while casual chat results in a high amount of noise. To mitigate this effect, we propose a technique for automatic detection of smalltalk based on the likelihood of spoken words being present in everyday dialogs. While visually recognizing driver's intention by learning from natural dialogs only is a challenging task, *learning from less but better data* via our smalltalk refinement consistently improves performance.**

I. Introduction and Related Work

For most driver observation frameworks, manually annotated datasets are regarded as the default starting point for training models [1]–[16]. While most road fatalities are caused by inappropriate vehicle maneuvers due to by human error [17], visual systems for detecting such events are often restricted to either a static sensor setup [1] or detecting a single maneuver type [11]. Applications of such methods at a large-scale are mostly hindered through expensive data collection requiring accurate temporal localization of the events. *Can we skip expensive domain specific annotations?*

The intersection of vision and language allows us to leverage weak annotations inherently present in social media, as multimodal posts hold an enormous amount of information about concepts of our daily life. One example of such content are mock driving exams often publicized on YouTube by driving schools in order to give advice to a wider audience. Such tests build on active student-teacher interaction, where the instructor gives verbal directions about the next action to take (Figure 1). A driving exam usually starts with casual talk and the teacher explaining the procedure, followed by the driving session, where the student verbally directed about the next action (*e.g. turn right at the traffic lights*) and concluded by further everyday conversation and feedback.

We view conversations happening during driving sessions as an unprecedented opportunity for connecting speech to changes in human action and present the first framework

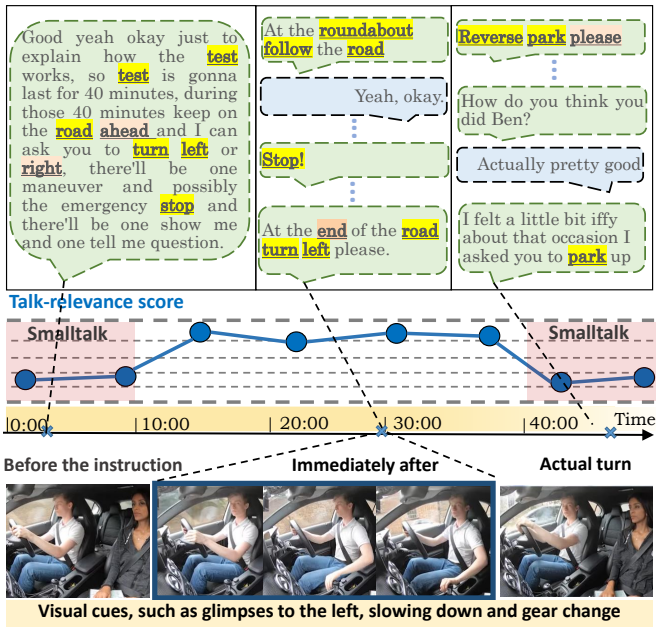

Fig. 1: **Example of a possible driving exam dialog.** The teacher verbally directs the student driver, triggering an imminent behavioral reaction, such as glimpses to the left. We leverage such conversations to learn common visual cues preceding the maneuver. To mitigate adverse effects of unrelated casual talk, we propose a technique for rating segment relevance based on the likelihood of spoken words being present in an everyday dialog compared to our setting.

for anticipating driver intent without a single manually labeled example. We collect a multimodal dataset of mock driving exams by querying YouTube and subsequently use conversation transcripts to learn charachteristic visual cues preceding certain maneuvers. While large portion of such recordings comprise the teacher giving concise directions, a challenge arises from the unrestricted nature of such dialogs, as casual talk is also common. To mitigate adverse effects of such segments without additional annotations, we propose a technique to detect such smalltalk by relating the likelihood of words spoken during driving conversation to their frequency in everyday speech, obtained from a movie dialogue corpus [18]. We empirically analyze the dialogs and use frequency and domain-distinctiveness of used terms

to derive seven maneuvers which we aim to predict by learning from ten second videos immediately following the request. As our visual model, we adopt multiple video classification architectures, which we train using ten second videos immediately following the request. While our experiments reveal that visually recognizing human intent through dialog supervision is a challenging task, all evaluated models surpass the random baseline by a large margin, while *learning from less but better data* with our smalltalk refinement consistently leads to a better recognition rate.

Context of previous work. While recent works use social media content as weak labels [19]–[23], our dataset offers a distinct setting where a *dialog* between two people has an imminent effect on *future actions*. Besides, we are the first to use *weak dialog supervision* without any manual annotations for driver maneuver prediction, resulting in our setting being less restrictive than previous work bound by the cost of manual annotation, although our examples are noisier. For example, [1], [24], [25] predict lane changes and turns, but consider a fixed camera view, while others focus on a single maneuver or use a simulated environment [11], [26]. Due to the free nature of web content, our dataset covers diverse views, people (almost all recordings have different drivers) and situations.While we focus on seven common events in our evaluation, new maneuvers can be added by issuing suitable dialog queries. While multiple important works target a related problem of maneuver prediction through vehicle signals (*e.g.* steering wheel angle) or a road camera view [13], [27] this is out-of-scope of this work, as we address vehicle maneuver prediction *based on driver observation* [1], [11], [11], [24]–[26], although extension of our framework with cameras facing the road is a potential future direction. While few recent driver observation works go beyond fully annotated training data *e.g.* considering the possibility of unknown behaviours occuring at test-time [28], [29], zero-shot learning [30] or domain shifts [31], [32], previous research does not consider dialog-guided visual recognition inside the car. The works most similar to ours are the ones emerging from the recent *Talk2Car* benchmark [33]–[35], which extends an autonomous driving dataset with language annotations "commanding the vehicle". While Deruyttere *et al.* [33] also consider linking human language to driving behaviour, they target the *outside* view while we focus on learning driver-related visual cues captured *inside* the vehicle and use the naturally occuring driving exam dialogs as supervision, while [33] leverage Amazon Mechanical Turk to label the recordings with textual commands. Recently, Nagrani *et al.* [36] pursued the idea of learning to predict future human action from natural dialogs obtained from movie screenplays, which is a very similar direction, although we target driving situations and learn from *exam* conversations.

To the best of our knowledge, this is the first work exploring two areas: 1) "webly"[1]-supervised learning for driver observation and 2) studying verbal teacher-student conversations during driving exams, which is not only a valuable supervision signal for training models (as done in our work), but could be used in the future to research the psychology of such road test conversations. Besides the applications inside the vehicle, our dataset of driving exam dialogs represents an interesting new avenue for research of speech and multimodality. Our environment is unique as the student-instructor dialogs trigger an immediate visual response, therefore, opening an excellent opportunity for connecting vision and language to actions.

[1]"Webly"-supervised learning: weakly supervised learning, where the data and *loose* annotations are automatically collected by crawling the Web [37].

II. Learning from Driving Exam Dialogs

A. Web Mining Mock Driving Exams

We aim to unveil the task of visually foreseeing future human actions by learning from naturally occurring dialogs and issue YouTube queries with terms such as "mock driving exam" or "road test" to collect the data. The only restriction we made is bypassing videos where the YouTube preview shot indicates that they obviously do not focus on humans, as we aim to study human behavior-only.

Our data covers both: 1) transcripts of conversations between the driver student and the examiner and 2) visual recordings inside the vehicle cabin.In 98 cases the transcripts were available through the YouTube API, while for the remaining recordings we used the *autosub*[2] library for automatic speech recognition. Although 120 sessions were initially collected, 14 were omitted as our smalltalk detection technique (described in Section II-B) indicated that they did not contain any relevant conversations (*e.g.* the teacher giving the student tips without any actual driving). The examples cover the driver and, in certain cases the driving instructor and have a variety of views due to the unrestricted nature of videos posted online. Since our main focus is the driver, we manually crop videos to the driver-view-only, which in the future could also be done automatically through *e.g.* a person detection model.

B. Detecting Smalltalk

Everyday conversations, such as exchanging pleasantries and giving feedback are common at the beginning and end of a driving exam. Such casual dialog, which may also occur in the middle, introduces additional noise and is presumably unfavourable for both, constructing a fair evaluation set and training the model [38]. While the structure of smalltalk is different from the driving conversations, they still may contain keywords of maneuvers we may want to recognize therefore adversely affecting the model. For example, if we want to recognize drivers searching for a parking spot, the feedback line "*I felt a bit iffy about that occasion I asked you to park*", which contains the word *park* is not helpful, as visuals do not match the mentioned action.

To meet this challenge without additional annotations, we propose a simple yet effective pre-processing technique for detecting smalltalk. Conceptually, our method comprises two

[2]github.com/agermanidis/autosub

Fig. 2: *What are people talking about during driving exams*? Domain-salient dialogs visualized, word size highlights term occurence frequency.

Fig. 3: Statistic of dialog lines containing maneuver commands by split. Colors: **train_refined**, **train_smalltalk**, **val**, **test**

Fig. 4: Proportion of domain-salient words for different relative positions in the exam; mean (blue points) and variance (yellow area).

steps: 1) computing *domain-salient words i.e.* terms which are distinctive for driving dialog, and 2) determining *domain-salient dialog segments* by estimating a score $s^\star(m,d)$ for every minute m of every dialog d based on the pace of domain-salient words inside the segment and its neighbour segments. The method is formalized in Algorithm 1 and is now explained in more detail.

Algorithm 1 Detecting Smalltalk

Input: $W_{d,m}$ – words spoken at minute m of a dialog d; V_{drive} – driving vocabulary; $p(w_i|drive)$ and $p(w_i|norm)$ – line occurrence probabilities for driving and normal settings (computed beforehand for all $w_i \in V_{drive}$); parameters α (salient word threshold); r (window range); ε (for num. stability) and $V^\star_{drive} := \emptyset$.
Output: region-saliency scores $s^\star$ for each exam minute
// Compute domain-salient word set $V^\star_{drive}$
for all $w_i \in V_{drive}$ **do**
  $f(w_i) := p(w_i|drive) \cdot \log\left(\frac{p(w_i|drive)}{p(w_i|norm)+\varepsilon}\right)$
  if $f(w_i) < \alpha$ **then** $V^\star_{drive} := V^\star_{drive} \cup \{w_i\}$
  end if
end for
// Compute domain-salient dialog regions
for all minutes m in all exam dialogs d **do**
  $s(m,d) := \frac{|W_{d,m} \cap V^\star_{drive}|}{|W_{d,m}|}$
end for
// Connect the neighboring regions via sliding window
for all minutes m in all exam dialogs d **do**
  $s^\star(m,d) := \sum_{i=m-r}^{m+r} s(m,d) \cdot \frac{1}{2r+1}$
end for

Preprocessing and word occurrence probabilities. First, we set all words to lower case, remove common English stop words and compute the vocabulary $\mathcal{V}_{drive}$ of all words that appear in the exam conversations ($|\mathcal{V}_{drive}| = 9492$). In a similar fashion, we set up $\mathcal{V}_{norm}$ with words spoken in movie conversations obtained from $5K$ lines randomly sampled from a movie dialog corpus [18] as our reference for everyday speech. We therefore explicitly distinguish between the *normal* and the *driving dialog corpora*. For each word $w_i \in \mathcal{V}_{drive}$ we compute $p(w_i|norm)$ and $p(w_i|drive)$, which are probability estimates of each word being present in a *normal* dialog sentence or *driving* corpora. To achieve this, we treat *dialog lines as documents* and compute the *document-term matrix* to model such probabilities.

Computing the domain-saliency score of words. Next, we aim to quantify the *domain-saliency* of a word, *i.e.* , how specific a certain term is to our *driving domain*. When designing such metric, we aim for words which (1) are *common* in driving conversations and (2) are *specific* to this context. In information retrieval systems, these requirements are often met by the *Term Frequency – Inverse Document Frequency (tf–idf)* technique [39], [40] which quantifies how important a word is to a document in a collection of documents by weighting its frequency in a document by inverse document frequency of the word across a set of documents. We cannot use *tf-idf* as-is due to a slightly different setting (we want to detect domain-relevant lines and therefore compute word occurrences in a dialog line, treating lines as documents, while having *two* distinct corpora). Inspired by *tf-idf* we derive a very similar metric to obtain the domain-saliency of a word in our setting. We compute the domain-saliency score $f(w_i)$ as the likelihood of w_i occurring in a driving conversation sentence weighted by the logarithm of ratio by which the probability estimates increases if the conversation has driving context:

$$f(w_i) = \underbrace{p(w_i|drive)}_{\text{How likely during driving?}} \cdot \log \underbrace{\left(\frac{p(w_i|drive)}{p(w_i|norm)+\varepsilon}\right)}_{\text{How specific to driving?}}, \quad (1)$$

where ε is an arbitrary small constant added for numerical stability. Note, that $p(w_i|drive)$ is the precomputed probability estimate of the word being present in a *driving* dialog sentence and $\frac{p(w_i|drive)}{p(w_i|norm)}$ signals the significance of the word in the driving domain, as it depicts by how much the occurrence probability increases if the conversation happened during driving. An additional logarithm transformation is used to "soften" the effect of the probability estimates increase term $\frac{p(w_i|drive)}{p(w_i|norm)}$ as it otherwise would become very high for words which have never been observed in the normal corpus.

We then refer to a word as *domain-salient* if its relevance score surpasses a threshold α (set to 0.002 producing $\sim$ 150 domain-salient words). Examples of such terms are *roundabout* and *exit*, which occur in *driving* conversations 51 and 43 times more often.

Identifying smalltalk segments We choose *one minute* as the temporal resolution for the smalltalk detection, and

then we rate these one minute segments as the *proportion of the used domain-salient* words. Then, we connect the neighbouring regions via sliding window smoothing (window range $r = 5$). We view a segment as smalltalk, if $s^{\star}(m,d) < 0.05$, *i.e*. the portion of domain salient words is less than 5% (after the region smoothing).

In Figure 5, we illustrate an example of the detected smalltalk regions for one session, in relation to the speech pace and the percentage of domain-salient words.

C. Dialog Analysis and Split Statistics

Using driver exam conversations as weak annotations allows us to query the specific maneuver type. For example, if we want to recognize searching for a parking spot or exiting the highway, we might look for terms such as *park* or *exit* in the dialogs. While further events can be added on-demand by issuing corresponding requests, we need to fix a category set for the evaluation. To achieve this, we took into consideration the terms with the highest domain-saliency score, as well as, looked into studies of maneuver impact on accident odds [41]. Finally, we inferred seven maneuvers: *stop*, *exit*, *park*, *turn right*, *turn left*, *straight* and *roundabout*.

Our dataset is split into *train*, *val*, and *test* sets with a 6:2:2 ratio of exams. As we always desire clean test data, *val* and *test* contain non-smalltalk dialog segments only. The *train* dialogs cover both, *train_smalltalk* and *train_refined*, which are regions that our approach marked as smalltalk or non-smalltalk. As our training data, we compare two options: all dialogs *train_refined* $\cup$ *train_smalltalk* and domain-salient *train_refined* only. Sample statistics by category and split, *i.e*. the number dialog lines containing the seven target commands, is provided in Figure 3.

The collected 106 recordings last 62.6 min on average. The mean speech pace of 71.9 words per minute (wpm) is significantly higher for the detected smalltalk (96.7 wpm) and lower for the instructional dialogs (29.6 wpm). Frequently used terms are illustrated in Figure 2: driving-related expressions (*i.e*. *road*, *turn*) overshadow the dialogs, while certain casual terms accompanying friendly request-response dialogs are also common (*e.g*. *please*, *thanks*). The average proportion of domain-salient words spoken in a dialog line (7.8% overall) is, unsurprisingly, higher for regions we estimate to be relevant (18%) and lower for smalltalk (1.8%). There is also a connection between the point in time and the speech relevance (Figure 4).

D. Visual Model

We use 10 second videos right after the teacher's request, to learn visual cues preceding the specific maneuvers. For our visual models, we implement three approaches based on spatiotemporal CNNs, initially developed for activity recognition: C3D [42], Inflated 3D ConvNet [43] and Pseudo3D ResNet [44]. The model weights are initialized using the Kinetics [43] dataset.

C3D C3D [42] is the first widely-used CNN leveraging 3D convolutions for action recognition. C3D consists of 8 convolutional layers ($3\times3\times3$ kernels) and 5 pooling layers ($2\times2\times2$) followed by two fully-connected layers. Besides being the first framework for generic spatiotemporal feature extraction from videos, it is compact and efficient through the small kernel sizes. It takes a 112×112 video snippet of 16 frames as input and produces a 4096-dimensional video feature followed by a fully-connected layer.

Inflated 3D ConvNet Inflated 3D ConvNet (I3D) [43] builds upon the Inception-v1 network by extending the 2D filters with an additional temporal dimension. The complete I3D network consists of 27 layers: three convolution layers at the beginning and one fully-connected layer at the end, four max-pooling layers at the beginning and one average pooling layer preceding the last fully-connected and nine inception modules which themselves are two layers deep. The input to I3D are 64 frames at 224×224 resolution.

Pseudo 3D ResNet Apart from such 2D-to-3D inflation, another way for reusing the available 2D CNNs on spatiotemporal video data is to first convolve them spatially (where pre-trained 2D models can be used), and then convolve the time dimension only. Following this paradigm, Pseudo 3D ResNet (P3D) [44] "mimics" 3D convolutions by combining a filter on the spatial domain (*i.e*. $3\times3\times1$) with one in the temporal dimension (*i.e*. $1\times1\times3$). P3D ResNet leverages residual connections due to improve the gradient flow, allowing a remarkable depth of 152 layers. P3D operates on 64 frame snippets with 160×160 pixel resolution.

We train our networks end-to-end using stochastic gradient descent (SGD) with momentum, employing early stopping on the validation set with the maximum number of epochs set to 200. All hyperparameters except for the learning rate and batch size are set as in the original papers presenting the models [42]–[44]. We use the batch size of 6 and we leverage grid search on the validation set to estimate the learning rate, dividing the current value by 10 in the range from 0.1 to 0.0001 (each trial conducted once). The learning rate is decreased by factor 0.2 after 50 and 100 epochs. The estimated initial learning rates are 0.001 fpr C3D and 0.01 for both, Pseudo 3D ResNet and Inflated 3D ConvNet. The models were trained on the Nvidia GeForce GTX 1080 Ti GPU with the training lasting $\sim$ 27 hours for Inflated 3D ConvNet, $\sim$ 17 hours for Pseudo 3D ResNet and $\sim$ 13 hours for C3D. As the classes are not equally represented in our dataset, we balance the dataset during training by oversampling underrepresented classes.

E. Experiments

We use balanced accuracy (mean accuracy over all classes) as our main metric and, additionally, report the unbalanced accuracy and the $F1$-score (harmonic mean of precision and recall) for the individual classes. To determine, whether it is better to learn from more but noisier data or from a smaller amount of relevant data, we consider two training settings: using complete training set dialogs or limiting them to non-smalltalk regions (*i.e*. skipping around 35% of the data where a maneuver term was present, but presumably in

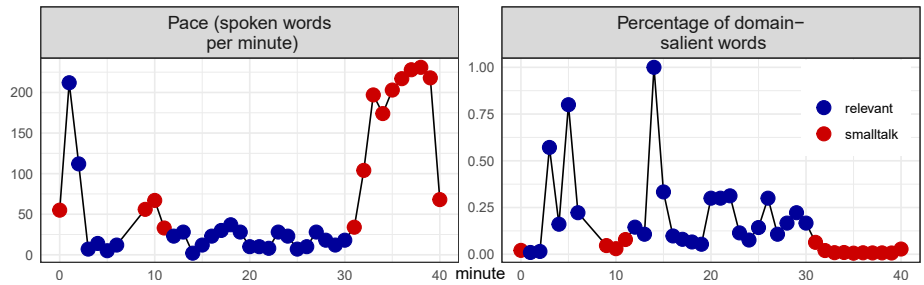

Fig. 5: Detected smalltalk regions (red) for one driving session example. The X-axis depicts the time (in minutes), while the Y-axis is the *speech pace* (words per minute) for the left graph and the *percentage of used domain-salient* words for the right graph. High pace (right graph) and rare usage of domain-salient words (left graph) are characteristic for smalltalk conversations behind the steering wheel.

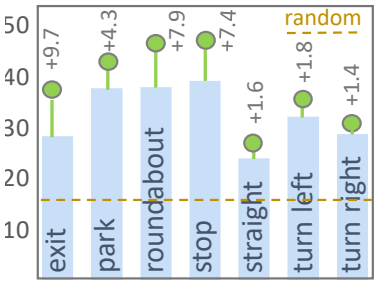

Fig. 6: Per-category F1 score for I3D with (green) and without (blue) our smalltalk refinement. Learning from *less but better* data through our refinement leads to better results.

Model Type	Smalltalk refined	Top1 Acc Bal.	Top1 Acc Unbal.
Random	–	33.33	33.33
CNNs without training set refinement			
C3D	no	47.62	44.02
Pseudo 3D Resnet	no	45.40	43.73
Inflated 3D Net	no	50.17	**55.43**
CNNs with training set smalltalk deletion			
C3D	yes	51.92	54.89
Pseudo 3D Resnet	yes	50.63	54.35
Inflated 3D Net	yes	**53.42**	**55.43**

TABLE I: Recognition results for the three-maneuver-setting (classes *straight*, *exit* and *stop*).

Model Type	Smalltalk refined	Top1 Acc Bal.	Top1 Acc Unbal.	Top3 Acc Bal.	Top3 Acc Unbal.
Random	–	14.29	14.29	42.86	42.86
CNNs without training set relevance refinement					
C3D	no	28.15	23.11	61.38	61.19
Pseudo 3D Resnet	no	22.43	20.32	56.87	56.93
Inflated 3D Net	no	32.76	34.79	64.18	69.34
CNNs with smalltalk deletion in the training set					
C3D	yes	31.09	30.78	64.61	67.15
Pseudo 3D Resnet	yes	31.68	33.21	61.8	67.64
Inflated 3D Net	yes	**36.05**	**39.66**	**65.64**	**70.56**

TABLE II: Results for all seven maneuvers. Smalltalk refinement improves while models and Inflated 3D Net performs the best.

a non-driving context). We therefore learn model parameters on either *train_refined* or *train_refined* ∪ *train_smalltalk* (see Section II-C), select checkpoints and hyperparameters on *val*, and present final evaluation on *test*.

We report the results for a simpler setting with three maneuvers in Table I, then move to a harder task with seven distinct events in Table II and, finally, examine the performance for individual classes in Figure 6. While the Inflated 3D Net is a clear frontrunner (53.42% for three and 36.05% for seven categories), learning from less but better data through our smalltalk refinement improves the recognition for all architectures. While it is evident that learning visual cues of the intended maneuvers guided only by exam dialogs is a hard task, all models outperform the random baseline by a large margin.

Limitations While our work makes a step towards economical training data collection for intelligent vehicles, we acknowledge, that the recognition quality obtained with such weakly supervised frameworks is far below models trained on clean and labelled training data [1]. Both training and test data is prompt to a large amount of noise even after the smalltalk refinement step. Although our models consistently outperform the random baseline, the recognition accuracy of $\sim 50\%$ for three and $\sim 35\%$ for seven distinct categories indicate strong need for improvement before the framework becomes a production-ready system. Instead, our approach should be viewed as a complement to fully supervised methods aimed at facilitating further research of economical data collection for driver observation. Combining our "webly"-supervised approach with smaller but fully annotated datasets [1] is an important future direction. Besides, the query outcomes influencing the training data are impacted by biases present, *e.g.* in the YouTube search algorithm [45].

III. Conclusion

We introduced the problem of weakly supervised driver intention prediction from videos by learning from driving exam conversations behind the steering wheel.To meet the challenge of noisy smalltalk conversations, we introduced a pre-processing technique for identifying and skipping such regions. We showed through experiments featuring different deep models, that such dialogs can be successfully used as guides for learning to foresee human driving intent.

Acknowledgements The research leading to this results was partially supported by the Competence Center Karlsruhe for AI Systems Engineering (CC-KING, https://www.ai-engineering.eu) sponsored by the Ministry of Economic Affairs, Labour and Housing Baden-Württemberg.

REFERENCES

[1] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3182–3190.

[2] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *Transactions on intelligent transportation systems*, vol. 15, no. 6, pp. 2368–2377, 2014.

[3] L. Xu and K. Fujimura, "Real-time driver activity recognition with random forests," in *International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 2014.

[4] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, "Real-time distracted driver posture classification," *NeurIPS Workshop on Machine Learning for Intelligent Transportation Systems*, 2018.

[5] C. Zhao, B. Zhang, J. He, and J. Lian, "Recognition of driving postures by contourlet transform and random forests," *IET Intelligent Transport Systems*, vol. 6, no. 2, pp. 161–168, 2012.

[6] C. Yan, F. Coenen, and B. Zhang, "Driving posture recognition by convolutional neural networks," *IET Computer Vision*, 2016.

[7] S. Martin, E. Ohn-Bar, A. Tawari, and M. M. Trivedi, "Understanding head and hand activities and coordination in naturalistic driving videos," in *Intelligent Vehicles Symposium*, 2014, pp. 884–889.

[8] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, "Head, eye, and hand patterns for driver activity recognition," in *International Conference on Pattern Recognition*, 2014, pp. 660–665.

[9] S. Y. Cheng and M. M. Trivedi, "Vision-based infotainment user determination by hand recognition for driver assistance," *Transactions on intelligent transportation systems*, vol. 11, no. 3, pp. 759–764, 2010.

[10] M. Martin*, A. Roitberg*, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen, "Drive&Act: A Multi-modal Dataset for Fine-grained Driver Behavior Recognition in Autonomous Vehicles," in *ICCV*. IEEE, October 2019.

[11] A. Doshi and M. Trivedi, "A comparative exploration of eye gaze and head motion cues for lane change intent prediction," in *2008 IEEE Intelligent Vehicles Symposium*. IEEE, 2008, pp. 49–54.

[12] P. Gebert, A. Roitberg, M. Haurilet, and R. Stiefelhagen, "End-to-end Prediction of Driver Intention using 3D Convolutional Neural Networks," in *Intelligent Vehicles Symposium (IV)*. IEEE, 2019.

[13] B. Morris, A. Doshi, and M. Trivedi, "Lane change intent prediction for driver assistance: On-road design and evaluation," in *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2011, pp. 895–901.

[14] A. Roitberg, M. Haurilet, S. Reiß, and R. Stiefelhagen, "Cnn-based driver activity understanding: Shedding light on deep spatiotemporal representations," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–6.

[15] J. D. Ortega, N. Kose, P. Cañas, M.-A. Chao, A. Unnervik, M. Nieto, O. Otaegui, and L. Salgado, "Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *European Conference on Computer Vision*. Springer, 2020, pp. 387–405.

[16] Y. Rong, Z. Akata, and E. Kasneci, "Driver intention anticipation based on in-cabin and driving scene monitoring," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–8.

[17] S. Reynolds, M. Tranter, P. Baden, D. Mais, A. Dhani, E. Wolch, and A. Bhagat, "Reported Road Casualties Great Britain: 2016 Annual Report," UK Department for Transport, Tech. Rep. September, 2017.

[18] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs." in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

[19] L. Breitfeller, E. Ahn, D. Jurgens, and Y. Tsvetkov, "Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts," in *International Conference on Empirical Methods in Natural Language Processing(EMNLP)*.

[20] D. Chinnappa, S. Murugan, and E. Blanco, "Extracting possessions from social media: Images complement language," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 663–672.

[21] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2630–2640.

[22] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3537–3545.

[23] R. Gomez, L. Gomez, J. Gibert, and D. Karatzas, "Learning to learn from web data through deep semantic embeddings," in *European Conference on Computer Vision (ECCV) Workshops*.

[24] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3118–3125.

[25] S. Martin, S. Vora, K. Yuen, and M. M. Trivedi, "Dynamics of driver's gaze: Explorations in behavior modeling and maneuver prediction," *IEEE Transactions on Intelligent Vehicles*, vol. 3.

[26] F. Yan, M. Eilers, L. Weber, and M. Baumann, "Investigating initial driver intention on overtaking on rural roads," in *Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 4354–4359.

[27] S. Lee, M. Q. Khan, and M. N. Husen, "Continuous car driving intent detection using structural pattern recognition," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[28] A. Roitberg, C. Ma, M. Haurilet, and R. Stiefelhagen, "Open set driver activity recognition," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1048–1053.

[29] O. Kopuklu, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 91–100.

[30] S. Reiß, A. Roitberg, M. Haurilet, and R. Stiefelhagen, "Activity-aware attributes for zero-shot driver behavior recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 902–903.

[31] ——, "Deep classification-driven domain adaptation for cross-modal driver behavior recognition," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1042–1047.

[32] J. S. Katrolia, B. Mirbach, A. El-Sherif, H. Feld, J. Rambach, and D. Stricker, "Ticam: A time-of-flight in-car cabin monitoring dataset," 2021.

[33] T. Deruyttere, S. Vandenhende, D. Grujicic, L. Van Gool, and M.-F. Moens, "Talk2car: Taking control of your self-driving car," *arXiv preprint arXiv:1909.10838*.

[34] H. Dai, S. Luo, Y. Ding, and L. Shao, "Commands for autonomous vehicles by progressively stacking visual-linguistic representations," in *European Conference on Computer Vision*. Springer, 2020, pp. 27–32.

[35] V. Mittal, "Attngrounder: Talking to cars with attention," in *European Conference on Computer Vision*. Springer, 2020, pp. 62–73.

[36] A. Nagrani, C. Sun, D. Ross, R. Sukthankar, C. Schmid, and A. Zisserman, "Speech2action: Cross-modal supervision for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 317–10 326.

[37] X. Chen and A. Gupta, "Webly supervised learning of convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1431–1439.

[38] H. Khayrallah and P. Koehn, "On the impact of various types of noise on neural machine translation," in *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, 2018, pp. 74–83.

[39] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information science*, vol. 27, no. 3, pp. 129–146, 1976.

[40] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of documentation*, 2004.

[41] S. Singh, "Critical reasons for crashes investigated in the national motor vehicle crash causation survey," Stats (National Highway Traffic Safety Administration), Rep. No. DOT HS 812 115., Tech. Rep., 2015.

[42] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *International Conference on Computer Vision*, 2015, pp. 4489–4497.

[43] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[44] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *Proceedings of the International Conference on Computer Vision*, 2017, pp. 5533–5541.

[45] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, *et al.*, "The youtube video recommendation system," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 293–296.